


N° d'ordre NNT : 2017-LYSET-008

## THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

**l'Ecole Nationale des Travaux Publics de l'Etat**

**École Doctorale** N° 162

**MEGA (Mécanique, Energétique, Génie Civil et Acoustique)**

**Spécialité / discipline de doctorat : Génie Civil**

soumise en vue d'une soutenance publique le 09 Octobre 2017, par :

**Nicole SCHIPER**

---

# Échantillonnage des données de trafic pour l'estimation de la pollution atmosphérique aux différentes échelles urbaines

---

Devant le jury composé de :

| | | |
|---|---|---|
| LATIFA OUKHELLOU, | Directrice de recherche (Université de Marne-la-Vallée) | *Rapporteur* |
| NIKOLAOS GEROLIMINIS, | Associate Professor (École polytechnique de Lausanne) | *Rapporteur* |
| CHRISTINE SOLNON, | Professeur (Université de Lyon) | *Présidente* |
| DELPHINE LEJRI, | Ingénieur des TPE, Docteur (Université de Lyon) | *Encadrante* |
| LUDOVIC LECLERCQ, | Directeur de Recherche (Université de Lyon) | *Directeur de Thèse* |

Thèse préparée au LICIT (Laboratoire d'Ingéniérie Circulation Transport)




N° d'ordre NNT : 2017-LYSET-008

**THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON**

Opérée au sein de

**l'Ecole Nationale des Travaux Publics de l'Etat**

**École Doctorale** N° 162

**MEGA (Mécanique, Energétique, Génie Civil et Acoustique)**

**Spécialité / discipline de doctorat : Génie Civil**

proposée pour soutenance publique le 09 Octobre 2017, par :

**Nicole SCHIPER**

---

# Traffic data sampling for air pollution estimation at different urban scales

---

Devant le jury composé de :

| | | |
|---|---|---|
| LATIFA OUKHELLOU, | Directrice de recherche (Université de Marne-la-Vallée) | *Rapporteur* |
| NIKOLAOS GEROLIMINIS, | Associate Professor (École polytechnique de Lausanne) | *Rapporteur* |
| CHRISTINE SOLNON, | Professeur (Université de Lyon) | *Présidente* |
| DELPHINE LEJRI, | Ingénieur des TPE, Docteur (Université de Lyon) | *Encadrante* |
| LUDOVIC LECLERCQ, | Directeur de Recherche (Université de Lyon) | *Directeur de Thèse* |

Thèse préparée au LICIT (Laboratoire d'Ingéniérie Circulation Transport)

**Nicole SCHIPER**
*Échantillonnage des données de trafic pour l'estimation de la pollution atmosphérique aux différentes échelles urbaines*
09 Octobre 2017
Rapporteurs : Latifa OUKHELLOU et Nikolaos GEROLIMINIS
Directeur de Thèse : Ludovic LECLERCQ ; Encadrante : Delphine LEJRI





*Thèse préparée au laboratoire*
**LICIT**
Laboratoire Ingéniérie CIrculation Transports
UMR T_E IFSTTAR-ENTPE
Université de Lyon

IFSTTAR
COSYS/LICIT
25 Avenue François Miterrand
Case 24
Cité des mobilités
69675 Bron Cedex, France

ENTPE
LICIT
Rue Maurice Audin
69518 Vaulx-en-Velin Cedex, France

# Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor HDR. Ludovic Leclercq for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Also, my sincere thanks goes to Dr. Delphine Lejri for supporting me to successfully completing my thesis, and for her kind assistance, insightful comments and knowledge that added considerably to my academic experience. I really appreciate her kindness assistance and support. My research would not have succeeded without her help.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Nikolaos Geroliminis, Dr. Latifa Oukhellou, and Prof. Christine Solnon, for their encouragement, insightful comments, and hard questions.

I thank my fellow labmates in AMMET: Christine Buisson, Nicolas Chiabaut, Cécile Becarie, Charlotte Duruisseau, Jean Krug, Anastasia Founta, Ségio Batista, Arthur Burianne, Johanna Cattin, Humberto Ramirez, Guilhem Mariotte et Chaunlin Zhao, Daniel Villegas et Etienne Hans for the stimulating discussions and for all the fun we have had in the last three years. A special acknowledgement goes to my office mate of many years: Clélia Lopez. She is a true friend ever since we began our thesis. Clélia is an amazing person in too many ways. I would like to thank the all people in the MOMI group, in special to Anne-Christine Demanny for her help and Nour-Eddin El-Faouzi for his advises.

A very special gratitude goes out to all down at IFSTTAR for helping and providing the funding the PhD research.

I am also grateful to the following ENTPE staff: Sonia Cenille, Francette Pignard, Luc Delattre and all the PhD students at ENTPE for their unfailing support and assistance. What a cracking place to work! With a special mention to Pierre Gayte to introduce me in the fantastic world of PhD students and their responsibilities inside of ENTPE' administration.

Getting through my thesis required more than academic support, and I have many, many people to thank for listening to and, at times, having to tolerate me over the past three years. I cannot begin to express my gratitude and appreciation for their friendship. A very special thanks to all my friends in Brazil, for their personal and professional support during the time I spent abroad. And cannot forget all the support from the Britto's family and most special Rafael Britto, every time I was ready to quit, you did not let me, and I am forever grateful.

For many memorable evenings out and in, I must thank everyone above as well as: Romain Lictevout, Luisana Gutierrez, Luc Fontbonne, Jeremy Humel and Rosana Gutierrez for all the support and fun. In particular, I am grateful to Dr. Diego Ramirez, for his unconditional support at work and

at life, he is a true friend.

Last but not the least, I would like to thank my family for supporting me spiritually throughout my life.

Thanks for all your encouragement!

# Résumé

La circulation routière est une source majeure de pollution atmosphérique dans les zones urbaines. Les décideurs insistent pour qu'on leur propose de nouvelles solutions, y compris de nouvelles stratégies de management qui pourraient directement faire baisser les émissions de polluants. Pour évaluer les performances de ces stratégies, le calcul des émissions de pollution devrait tenir compte de la dynamique spatiale et temporelle du trafic. L'utilisation de capteurs traditionnels sur route (par exemple, capteurs inductifs ou boucles de comptage) pour collecter des données en temps réel est nécessaire mais pas suffisante en raison de leur coût de mise en œuvre très élevé. Le fait que de telles technologies, pour des raisons pratiques, ne fournissent que des informations locales est un inconvénient. Certaines méthodes devraient ensuite être appliquées pour étendre cette information locale à une grande échelle. Ces méthodes souffrent actuellement des limites suivantes : (i) la relation entre les données manquantes et la précision de l'estimation ne peut être facilement déterminée et (ii) les calculs à grande échelle sont énormément coûteux, principalement lorsque les phénomènes de congestion sont considérés. Compte tenu d'une simulation microscopique du trafic couplée à un modèle d'émission, une approche innovante de ce problème est mise en œuvre. Elle consiste à appliquer des techniques de sélection statistique qui permettent d'identifier les emplacements les plus pertinents pour estimer les émissions des véhicules du réseau à différentes échelles spatiales et temporelles. Ce travail explore l'utilisation de méthodes statistiques intelligentes et naïves, comme outil pour sélectionner l'information la plus pertinente sur le trafic et les émissions sur un réseau afin de déterminer les valeurs totales à plusieurs échelles. Ce travail met également en évidence quelques précautions à prendre en compte quand on calcul les émissions à large échelle à partir des données trafic et d'un modèle d'émission. L'utilisation des facteurs d'émission COPERT IV à différentes échelles spatio-temporelles induit un biais en fonction des conditions de circulation par rapport à l'échelle d'origine (cycles de conduite). Ce biais observé sur nos simulations a été quantifié en fonction des indicateurs de trafic (vitesse moyenne). Il a également été démontré qu'il avait une double origine : la convexité des fonctions d'émission et la covariance des variables de trafic.

**Mots-Clés :** *Émissions de véhicules*, *Échantillonage de données de trafic*, *Agrégation spatio-temporelle*, *Échelles de réseau*, *Modèle d'émission.*

## Abstract

Road traffic is a major source of air pollution in urban areas. Policy makers are pushing for different solutions including new traffic management strategies that can directly lower pollutants emissions. To assess the performances of such strategies, the calculation of pollution emission should consider spatial and temporal dynamic of the traffic. The use of traditional on-road sensors (e.g. inductive sensors) for collecting real-time data is necessary but not sufficient because of their expensive cost of implementation. It is also a disadvantage that such technologies, for practical reasons, only provide local information. Some methods should then be applied to expand this local information to large spatial extent. These methods currently suffer from the following limitations: (i) the relationship between missing data and the estimation accuracy, both cannot be easily determined and (ii) the calculations on large area is computationally expensive in particular when time evolution is considered. Given a dynamic traffic simulation coupled with an emission model, a novel approach to this problem is taken by applying selection techniques that can identify the most relevant locations to estimate the network vehicle emissions in various spatial and temporal scales. This work explores the use of different statistical methods both naïve and smart, as tools for selecting the most relevant traffic and emission information on a network to determine the total values at any scale. This work also highlights some cautions when such traffic-emission coupled method is used to quantify emissions due the traffic. Using the COPERT IV emission functions at various spatial-temporal scales induces a bias depending on traffic conditions, in comparison to the original scale (driving cycles). This bias observed in our simulations, has been quantified in function of traffic indicators (mean speed). It also has been demonstrated to have a double origin: the emission functions' convexity and the traffic variables covariance.



**Keywords :** *vehicle emissions*, *traffic data selection*, *spatial-temporal aggregation*, *network scales*, *emission model*.

# Table of Contents

**List of Figures** **6**

**List of Tables** **8**

**1 Estimating road emissions from traffic models** **19**
1.1 Description of model . . . 19
1.1.1 Traffic models . . . 19
1.1.2 Emission models . . . 23
1.2 Coupling traffic and emission models . . . 29
1.3 General objectives of the thesis work . . . 30

**2 Descriptive analysis of a network transportation - Paris case** **33**
2.1 Traffic simulations . . . 33
2.1.1 Simulation environment . . . 33
2.1.2 Data Generation . . . 35
2.2 Traffic and pollutant emission variables . . . 37
2.3 Influence of the variable definitions . . . 38
2.3.1 Comparison between spatial and punctual sensors . . . 38
2.3.2 Pollutant Emissions . . . 43
2.3.3 Conclusions about the variable definitions . . . 47
2.3.4 Further spatial variable analysis . . . 48
2.4 Conclusion of Chapter 2 . . . 56

**3 Sampling link selection based on the LASSO method** **59**
3.1 LASSO . . . 60
3.1.1 The idea underlying the method . . . 60
3.2 Datasets . . . 63
3.2.1 Static/static dataset . . . 63
3.2.2 Dynamic/dynamic dataset . . . 64
3.2.3 Static/dynamic dataset . . . 65
3.3 LASSO method applied to the three different datasets . . . 65
3.3.1 LASSO applied to the static/static dataset . . . 65

3.3.2 LASSO applied to static/dynamic datasets . . . 85
3.3.3 LASSO applied to dynamic/dynamic datasets . . . 108
3.3.4 Comparison between datasets . . . 124
3.4 Conclusion of the chapter . . . 129

**4 Model selection from other statistical methods 131**
4.1 Random Selection . . . 132
4.1.1 Static/static dataset . . . 132
4.1.2 Static/dynamic dataset . . . 141
4.2 Ranked links . . . 147
4.2.1 Static/static datasets . . . 148
4.2.2 Static/dynamic datasets . . . 149
4.3 Stepwise selection . . . 152
4.3.1 Static/static datasets . . . 154
4.3.2 Static/dynamic datasets . . . 155
4.3.3 Conclusion . . . 157
4.4 Comparison of all the methods . . . 158
4.4.1 Static/static dataset . . . 161
4.4.2 Static/dynamic dataset . . . 163
4.5 Case of application . . . 165
4.6 Conclusion of the chapter . . . 167

**5 Effect of traffic data aggregation on emission estimations 169**
5.1 Local versus global emission calculations . . . 170
5.1.1 Data input . . . 170
5.1.2 Emission quantification . . . 171
5.2 The analysis of the biases . . . 173
5.2.1 Driving cycles . . . 173
5.2.2 Emission gap as a function of data aggregation . . . 175
5.3 Conclusion of the chapter . . . 181
5.4 The results obtained . . . 184
5.5 Perspectives . . . 186

**Appendices 197**

**A Correlations between variables by period of time 201**

**B Static/Static dataset appendix 207**
B.1 Stability of LASSO selection . . . 207
B.2 Model size and the error distribution for static/static datasets . . . 210
B.2.1 Traveled distance . . . 210
B.2.2 Spatial mean speed . . . 210
B.3 Influence of route choice on the selection . . . 213

**C Static/Dynamic dataset appendix** **217**
C.1 Models proposed by LASSO . . . 217
C.2 Model size and error distribution . . . 219
C.3 The robustness of the models . . . 223
C.4 The best model for each variable . . . 224
C.5 The influence of route choice on the selection . . . 226

**D Dynamic/Dynamic dataset appendix** **231**
D.1 Model size and error distribution . . . 231

**E Comparison between datasets** **235**

**F Network partitioning for the static/static dataset** **239**
F.1 $NO_x$ network partitioning . . . 239

**G Lottery method applied to the static/dynamic dataset** **243**
G.1 $CO_2$ outliers . . . 243
G.2 $NO_x$ outliers . . . 244

**H Network partitioning for the static/dynamic dataset** **247**
H.1 $CO_2$ emissions . . . 247
H.2 $NO_x$ emissions . . . 248

**I Stepwise selection** **251**

**J Comparison of all methods** **253**
J.1 Static/dynamic datasets . . . 253

# List of Figures


1.1 Level of detail of the representation of vehicle flows (Zehe, 2015) . . . . . 22
1.2 Representation of traffic and emission models (Rocha, 2013) . . . . . 29

2.1 OD flow coefficient for each time period and the entries where the demand values are applied.(Villegas et al., 2013) . . . . . 34
2.2 The network profile made in Symuvia with the traffic light positions. The major axes of the network are indicated with colors. The demand values are shown in each entry, and they correspond to the traffic peak hours of the demand values.(Villegas et al., 2013) 35
2.3 Intensity of flow directions considered in the morning and the evening time. (Villegas et al., 2013) . . . . . 36
2.4 OD flow coefficients for each period of time in Boulevard Saint Germain. . . . . 37
2.5 The traffic variables recovered from each link and time period of the network and their respective emission calculations. . . . . 39
2.6 Network traveled distance comparison. . . . . 40
2.7 The error distributions of daily network traveled distance and their relative mean errors. 41
2.8 Network mean speed comparisons. . . . . 42
2.9 The error distribution comparison of daily network mean speed. . . . . 42
2.10 Emission factors determined by COPERT IV methodology for passenger cars considering 2015 fleet composition for hot exhaust emissions . . . . . 43
2.11 Pollutant emission comparison . . . . . 45
2.12 Comparison of network emissions determined by COPERT IV methodology. E1 represent emissions from spatial sensors (reference values); E2 are emissions calculated using traveled distances determined by static length and; E3 were calculated using dynamic link length to determine traveled distance. . . . . 46
2.13 Fitted emission-speeds curves. . . . . 46
2.14 Hypothesis comparison. . . . . 47
2.15 Percentage difference between the hypothesis. . . . . 48
2.16 Traffic variable evolutions through time periods: All simulations are represented at network level. . . . . 48
2.17 Pollutant emissions through time periods: All simulations are represented. . . . . 49
2.18 Distribution of daily network variable values from all simulations. . . . . 50
2.19 Outliers simulation ID by variable. . . . . 51

2.20 Correlation between variables at network level. . . . . 52
2.21 Average values of variables by link considering more than a year of data. . . . . 53
2.22 10% of the most representative links for each variable in the network. . . . . 54
2.23 Major axes highlighted in red and minor ones in blue. . . . . 54
2.24 Network partitioning by variable. . . . . 55
2.25 Histogram of variables by clusters. . . . . 56

3.1 Situation where $A$ is inside the zone where $\beta_2$ is exactly zero (Friedman et al., 2010). . 62
3.2 Estimated prediction error curves for the variables in static/static datasets. . . . . 66
3.3 Comparison between the predicted values of each model proposed by LASSO and the respective reference values. . . . . 68
3.4 Selected links by the $\lambda$ calculated with one standard error rule (1SE) and the mean relative absolute error of each variable. . . . . 69
3.5 Percentage of error distribution between the predicted variables values by the model built with selected links and the reference values of variables (Y). . . . . 70
3.6 The percentage error of models with fewer predictors than those of optimized lambdas. 72
3.7 The percentage errors of models with more predictors than optimized lambdas. . . . . 72
3.8 The percentage errors of models with fewer predictors than optimized lambdas. . . . . 73
3.9 The percentage errors of models with more predictors than optimized lambdas. . . . . 73
3.10 The error percentages of model variable selections used to estimate the total traveled distance. . . . . 77
3.11 Percentage error of model variable selection for spatial mean speed estimation. . . . . 78
3.12 The percentage errors of the model selections used to estimate network $CO_2$ emissions. 78
3.13 The percentage error of the model selections for estimating network $NO_x$ emissions. . 79
3.14 Pollutant emission densities obtained using the observation of the morning peak as the training set and the evening peak as the validation set. . . . . 81
3.15 Network mean speed results using scenario 1 (morning) as the training set and scenario 2 (evening) as the validation set. . . . . 82
3.16 Network mean speed results using scenario 2 (evening) as the training set and scenario 1 (morning) as the validation set. . . . . 82
3.17 Links selected for mean speed for the morning and evening cases (evening). . . . . 83
3.18 Estimated prediction error curves and their standard errors for the variables in static/dynamic datasets. . . . . 85
3.19 Comparison between the reference values and the predicted values and their associated errors. . . . . 88
3.20 Selected links determined by the $\lambda$ model calculated with one standard error rule and the absolute average percentage of errors given by the model using the selected links. . 89
3.21 Percentage error between the predicted variable values and the response values. . . . . 90
3.22 The traveled distance error distributions from lambda models fitted using LASSO and that have fewer predictors than the 1 SE lambda model (all plots without their respective outliers). . . . . 91
3.23 Error distributions on the traveled distance models. . . . . 91
3.24 The error distributions of spatial mean speed for all lambdas with models with fewer predictors than 1SE lambda. . . . . 92
3.25 Error distributions on the spatial mean speed models. . . . . 92

3.26 The error distributions from network $CO_2$ emission models fitted using LASSO and that have fewer predictors than the 1 SE lambda without outliers. . . . . 93
3.27 Error distributions on the network $CO_2$ emission models. . . . . 94
3.28 The error distributions from network $NO_x$ emission models fitted using LASSO and that have fewer predictors than the 1 SE lambda without outliers. . . . . 94
3.29 Error distributions on the network $NO_x$ emission models. . . . . 95
3.30 Error distributions from absolute error values of the 1SE lambda model. . . . . 96
3.31 Error distribution of models used to determine the network traveled distance. . . . . 98
3.32 Error distribution of models that estimate the spatial mean speed without outliers. . . 98
3.33 Error distributions of models used to estimate the network $CO_2$ emissions without their outliers. . . . . 99
3.34 Error distribution of models used to estimate the network $NO_x$ emissions without their outliers. . . . . 99
3.35 Error distributions from models built using morning data and validated by the evening data. . . . . 103
3.36 Density distributions for the network variables. . . . . 104
3.37 Links selected by the 1 SE lambda model in scenario 1 for the network variables. . . . 105
3.38 Error distribution of variables. . . . . 106
3.39 Density values for all variable models. . . . . 107
3.40 Estimated prediction error curves and their standard errors for the variables in dynamic/dynamic datasets. . . . . 108
3.41 Comparisons of models for each variable. . . . . 110
3.42 Traveled distance selection. . . . . 111
3.43 Spatial mean speed selection. . . . . 112
3.44 $CO_2$ emission selection. . . . . 112
3.45 The selected time periods and links for $NO_x$ emissions. . . . . 113
3.46 Error percentage between the predicted variable values using the 1SE lambda model and the original values of the variables. . . . . 113
3.47 Error distributions versus number of predictors to estimate network traveled distance. 114
3.48 Error distributions versus number of predictors to estimate spatial mean speed. . . . . 115
3.49 Error distributions versus number of predictors to estimate network $CO_2$ emissions. . 116
3.50 Error distributions versus number of predictors to estimate network $NO_x$ emissions. . 116
3.51 Error distributions for models used to estimate the network traveled distance and the number of selected links in each model. . . . . 120
3.52 Error distributions of models used to estimate the spatial mean speed and the number of selected links inside each model. . . . . 121
3.53 Error distributions for models used to estimate the network $CO_2$ emissions and the number of links selected for each model. . . . . 122
3.54 Error distributions for models used to estimate the network $NO_x$ emissions and the number of links selected for each model. . . . . 122
3.55 Comparison between both optimized models with static/static and dynamic/dynamic datasets. . . . . 125
3.56 Comparison between: the reference values, the LASSO models and the application of static/static models to static/dynamic datasets for each variable. . . . . 126

3.57 The error distribution by time period for network traveled distance using the SS model applied in SD data. . . . . 126
3.58 The error distribution by time period for network mean speed using the SS model applied to SD data. . . . . 127
3.59 Comparison between: the reference values, the LASSO models and the application of dynamic/static models in static/static datasets for each variable. . . . . 128
3.60 Error distribution calculated using the difference between reference values and the values from the application of dynamic/static models in static/static datasets for each variable. 129

4.1 The error distributions of the daily $CO_2$ emission estimations of each random sampling. 133
4.2 Comparison of the $CO_2$ estimations from random selection and the LASSO model. . . 134
4.3 The error distribution of the daily $NO_x$ emission estimation for each sample. . . . . 135
4.4 Comparison of the $NO_x$ estimations from the random selection and the LASSO model. 135
4.5 The total normalized variance and the connected cluster dissimilarity indicators of $CO_2$ and $NO_x$ missions. . . . . 137
4.6 The partitioning of the Paris network based on daily $CO_2$ emissions. . . . . 137
4.7 Histogram of clustering in the Paris network based on daily $CO_2$ emissions. . . . . 138
4.8 The 11 links selected randomly in both cases of partitioning. . . . . 139
4.9 Density and associated error for the estimated $CO_2$ emission values for the LASSO model and the models based on the random selection into clusters. . . . . 139
4.10 The 11 links selected randomly in both cases of partitioning. . . . . 140
4.11 Density and associated error for the estimated $NO_x$ emission values for the LASSO model and the models based on the random selection into clusters. . . . . 141
4.12 The $CO_2$ error distributions for each random sampling. . . . . 142
4.13 The $CO_2$ error distributions without their respective outliers for each random sampling. 142
4.14 Density and associated errors of estimated $CO_2$ emission values based on the lottery method. . . . . 143
4.15 The $NO_x$ error distributions without their respective outliers for each random sampling. 144
4.16 Density and associated errors from the estimated $NO_x$ emission values based on the lottery method. . . . . 144
4.17 77 links selected randomly in both cases of partitioning. . . . . 145
4.18 Density and associated errors for the estimated $CO_2$ emission values for the LASSO model and the models based on the random selection in clusters. . . . . 146
4.19 The 65 links selected randomly in both cases of partitioning. . . . . 147
4.20 Density and associated error for the estimated $NO_x$ emission values for the LASSO model and the models based on the random selection in clusters. . . . . 147
4.21 Most pollutant links of the network. . . . . 148
4.22 Comparison between predicted values. . . . . 149
4.23 Most pollutant links of the network. . . . . 150
4.24 Densities of emission values. . . . . 150
4.25 Error distributions from the predicted values. . . . . 151
4.26 Densities and error distributions from the predicted values by each model. . . . . 155
4.27 Links selected by the stepwise method in the static/static dataset. . . . . 156
4.28 Densities and error distributions from the predicted values by each model. . . . . 157
4.29 Links selected by the stepwise method in the static/dynamic dataset. . . . . 158

4.30 The mean absolute error for each random sampling and dataset. . . . . 159
4.31 The mean absolute error for each random sampling and dataset. . . . . 159
4.32 Comparison of the daily emissions values of emissions estimated by the proposed methods. 162
4.33 Error comparison between the estimated values of all models to estimate the daily $CO_2$ emissions. . . . . 162
4.34 Comparison of errors between the estimated values of all the models used to estimate the daily $NO_x$ emissions. . . . . 163
4.35 Comparison of the density of the daily emission values estimated by the methods proposed. 164
4.36 Comparison of errors between the estimated values of all the models used to estimate the daily $CO_2$ emissions. . . . . 164
4.37 Comparison of errors between the values of all the models used to estimate daily $NO_x$ emissions. . . . . 165
4.38 Links equipped with traffic sensors in the $6^{\text{th}}$ district of Paris. . . . . 166
4.39 Error distribution of the network emission estimation. . . . . 167

5.1 Density distributions of network emissions using the local and global approaches. . . . 172
5.2 Percentage of disparity between the daily emission estimations for each pollutant. . . . 172
5.3 Disparities on emission estimations between scales by time period. . . . . 173
5.4 Example of the modal driving cycle (NEDC) (Barlow et al., 2009). . . . . 174
5.5 Example of a transient driving cycle (Barlow et al., 2009). . . . . 174
5.6 Hypothetical driving cycle. . . . . 175
5.7 Comparisons between the weighted and arithmetic mean speeds by time period. . . . . 178
5.8 Disparities on emission factors by time period. . . . . 179
5.9 Percentage of disparity considering the convexity of the function. . . . . 180
5.10 Correlation between traveled distance and emission factor. . . . . 181

A.1 Variable correlations through periods of time . . . . . 205

B.1 Comparison between selections made for different training and validation sets in travel production. . . . . 207
B.2 Comparison between selections made for different training and validation sets for mean speed. . . . . 208
B.3 Comparison between selections made for different training and validation sets for $CO_2$ emissions. . . . . 208
B.4 Comparison between selections made for different training and validation sets for $NO_x$ emissions. . . . . 209
B.5 The percentage error of models calculated with different lambdas. . . . . 210
B.6 The percentage error of models between 1SE lambda and lambda. . . . . 211
B.7 The percentage error from models with more predictors than the optimized models. . 211
B.8 The percentage error of models with fewer predictors than optimized lambdas. . . . . 212
B.9 The percentage error of models with more predictors than the optimized lambdas. . . 212
B.10 Pollutant emission densities using the observation of the evening peak as the training set and the morning peak as the validation set. . . . . 213
B.11 Traveled distance densities for both cases of the study. . . . . 214

C.1 Error distribution of variables in the static/dynamic dataset. . . . . 218

C.2 The error distributions of models fitted using LASSO and that have fewer predictors than the 1 SE lambda model. . . . 219
C.3 The error distributions for all lambdas that have models with fewer predictors than 1SElambda. . . . 220
C.4 The error distributions from network $CO_2$ emission models fitted using LASSO with fewer predictors than 1 SElambda. . . . 221
C.5 The error distributions from network $NO_x$ emission models fitted using LASSO and with fewer predictors than 1SElambda. . . . 222
C.6 Error distribution from models that estimate the spatial mean speed. . . . 224
C.7 Error distribution of models used to determine the network $CO_2$ emissions. . . . 224
C.8 Error distribution of models used to estimate the network $NO_x$ emissions. . . . 225
C.9 LASSO Cross-validation of variables using morning data. . . . 226
C.10 LASSO Cross-validation of variables using evening data. . . . 227
C.11 Selected links by variable using the evening data. . . . 228

D.1 Error distributions versus number of predictors to estimate network traveled distance. 231
D.2 Error distributions versus number of predictors to estimate spatial mean speed. . . . 232
D.3 Error distributions versus number of predictors to estimate network $CO_2$ emissions. . 233

E.1 Error distribution by time period for network $CO_2$ emissions using the SS model applied to SD data. . . . 235
E.2 Error distribution by time period for network $NO_x$ emissions using the SS model applied to SD data. . . . 236

F.1 Network partitioning into 4 clusters based on $NO_x$ missions. . . . 239
F.2 Network partitioning into 6 clusters based on the $NO_x$ missions. . . . 240

G.1 The $CO_2$ error distributions. . . . 243
G.2 The $NO_x$ error estimations for all the random draws taking into account the outliers. 244
G.3 The $NO_x$ error distributions. . . . 244

H.1 $CO_2$ error distributions. . . . 247
H.2 $NO_x$ error distributions. . . . 248

I.1 Comparison of associated errors from the LASSO and the stepwise model. . . . 251

J.1 $CO_2$ error distributions. . . . 253
J.2 $NO_x$ error distributions. . . . 254

# List of Tables


2.1 Percentage of common links present in the variable rankings. . . . . 54

3.1 Ratio of common selected links between variables. . . . . 70
3.2 The average absolute error of the model established for one variable and applied to another. . . . . 74
3.3 The mean squared error of the model established with one variable applied to another. 75
3.4 The average percentage of absolute error from the linear regression model fitted to the merge or intersection between variables of the same type. . . . . 75
3.5 The mean square error from the linear regression model fitted to the merging or intersection between variables of the same nature. . . . . 76
3.6 The Bayesian information criterion for each model studied in static/static datasets. . . 80
3.7 Ratio of common selected links between variables. . . . . 88
3.8 Percentage error of a model applied to a variable. . . . . 96
3.9 Percentage of absolute error from a model applied in a variable. . . . . 97
3.10 Percentage of outliers in each model. . . . . 100
3.11 Mean square error of each model and variable. . . . . 101
3.12 The Bayesian information criterion for all models in static/dynamic dataset. . . . . 101
3.13 Common links selected in variables. . . . . 104
3.14 Percentage of selected time periods for each model and variable. . . . . 118
3.15 The average absolute error of the 1 SE lambda model selection on one variable applied on another. . . . . 119
3.16 The common selected links and time periods between variables. . . . . 119
3.17 The average percentage absolute error of the linear regression model fitted to the merge and intersection between variables of the same type. . . . . 120
3.18 Bayesian information criterion for all models studied in the dynamic/dynamic dataset. 123

4.1 The Bayesian Information Criterion for the network emission estimation models. . . . 161

C.1 The mean square error of all the models. . . . . 223
C.2 Common links between variables. . . . . 228

# Abbreviations

These abbreviations are introduced, in order alphabetic:

**ANR** ..... The French National Research Agency

**BIC** ..... Bayesian information criteria

**COPERT IV** .. Computer program to calculate emissions from road transport - version IV

$CO_2$ ..... Carbon dioxide

$CO$ ..... Carbon monoxide

**DTP** ..... Total traveled distance

**EF** ..... Emission factor

**IGN** ..... The national institut of geographic and forest information

**ITS** ..... Intelligent Transportation Systems

**LASSO** ..... Least absolute shrinkage and selection operator

$l_{dyn}$ ..... Dynamic link length

$l_{bc}$ ..... Static link length

**MFD** ..... Fundamental diagram of traffic flow

$NO_x$ ..... Nitrogen oxides

**OD** ..... Origin-destination flow

**1SE** ..... One standard error rule

$PM$ ..... Particulate matter

$VOCs$ ..... Volatile organic compounds

**VIT** ..... Spatial mean speed

**VKT** ..... Vehicle kilometer traveled

# Introduction

## Transportation and the environment: vehicle emissions

Pollutant emissions are a serious concern around the world and have grown at a relatively considerable rate over the last 25 years. A large number of industrial sectors have participated in this growth including energy, processing, agriculture, waste treatment, solvents and other products. (DOE, 2015) showed that transportation will be the only sector where emissions will increase instead of stabilizing or decreasing in the near future given current and predicted growth trends in transportation.

According to (EEA, 2017a), the number of passenger cars in the EU-28 area increased on average by 1.2% per year between 2000 and 2013, which represents from 415 to 490 cars per 1000 inhabitants. This growth has resulted from other important factors such as: (i) a decrease in the number of persons per household, (ii) an increase in the number of cars per family, and (iii) an increase in average travel distance due to less access to public transport. Regarding the composition of the European vehicle fleet, the number of diesel cars increased from 27% to 38% for the same period in European Union (EU), moreover the percentage of diesel cars in France is particularly high and can reach 68% of the fleet. When it comes to pollutants emissions, light vehicles are responsible for approximately 85% of total road emissions (IPCC, 2013). Some cities of the EU present a considerable increase in nitrogen oxides ($NO_2$) and particulate matter ($PM_{2.5}$ and $PM_{10}$) measured close to traffic. The source of this pollution mainly diesel cars and is strongly related to poor quality air. The World Health Organization (WHO) set air quality guidelines to protect the human population in the EU-28 area and concluded that more than 90% of the population is exposed to one pollutant of at least dangerous levels health. Between 2006 and 2014, the urban population of the EU-28 was exposed to concentrations in excess of the target limits set by (WHO, 2013) and (EU, 2013): (i) for fine particulate matter ($PM_{2,5}$) 8-17% of the urban population was exposed to the EU target values and 85-97% was exposed to concentrations above the WHO guideline values; (ii) for particulate matter ($PM_{10}$) the respective exposures were 16-42% for the EU limit values and 50-92% for the WHO guidelines; (iii) for nitrogen dioxide, estimates were 7-31% for both limit values.

Considering greenhouse gas emissions (GHG) from transport, carbon dioxide ($CO_2$), methane ($CH_4$) and nitrous oxides ($N_2O$) emissions increased by 2% in 2015 in comparison to 2014 (EEA, 2017b) and by more than 19% versus the levels of 1990. In the same year, road transport was responsible for almost 73% of total greenhouses gas emissions from transport, and 44% of these emissions came from passengers cars while up to 18% were from heavy-duty vehicles. Considering climate changes, carbon dioxide ($CO_2$) emitted by traffic sources is considered one of the greenhouse

gases having the greatest impact on climate change (IPCC, 2013) and it represents an economic, social and environmental threat.

Taking into account all these concerns, many policies have been implemented to reduce the exposure of the population and thus increase air quality. Local and regional management plans have been drawn up to improve air quality, including initiatives such as low-emission zones in cities. At a larger scale, changes in laws and international political agreements are crucial strategies aimed at mitigate air pollution. In order to prevent the impacts of climate change, some countries have agreed to cooperate in view to limiting the increase of the global temperature and the resulting climate changes by signing the United Nations Framework Convention on Climate Change (UNFCCC). The aim of the UNFCCC is to require its signatories to maintain precise and regular inventories of GHG to prevent dangerous anthropic interference in the climate system (UNFCCC, 2014). Also at the international level, the Kyoto protocol is the main instrument for mitigating GHG emissions. It was signed in 1997 and set objectives to reduce emissions in the signatory country. The first action period of the Kyoto protocol started in 2008 and ended in 2012 while the second started in 2013 and will end in 2020. In the meantime, the EU decided its own climate change mitigation objective for 2020 by reducing emission levels by 40% by 2030 and by 80% by 2050 compared to 1990 levels (UNFCCC, 2012).

To reach the objectives given by the laws and agreements, actions have been taken at the local scale. Over the last decade, tighter standards have been introduced to increase the efficiency of these actions. Some examples of emission mitigation actions in European countries are: fitting catalytic diesel particulate filters to older vehicles and older buses to reduce $NO_x$ emission in urban areas (Carslaw and Beevers, 2005); the Low Emission Zone implemented in 2008 in London restricts the entry of the most polluting Heavy Good Vehicle (HGVs) in strategic areas (TfL, 2014); the roll out plan for new hybrid and low-emission vehicles (EURO IV); the air quality strategy also implemented in London in 2010 and used as a model for other European cities (GLA, 2010). Other alternatives have been implemented such as the use of renewable fuels in transportation and the incentive to increase the share of alternative-fuel vehicles in the total fleet, with the objective of reaching 10% of renewable energy in transport by 2020 in comparison to 5.4% in 2013 (EEA, 2016).

Nowadays, there is concern that these strategies have not performed as foreseen or that they are not strong enough to decrease emissions. (Font and Fuler, 2016) studied the impact of policies to reduce traffic-related emissions in London for a period of 5 years, and concluded that despite the reduction in the number of cars and taxis, the levels of pollutants, particularly carbon dioxide, did not decrease accordingly. Mobility management in large and developing cities is an important means of changing the way vehicles are used, in order to increase the capacity and efficiency of the transport system and thus reduce vehicle emissions. It also helps to improve traffic flows, therefore reducing congestion and thus emissions. The challenge is therefore to provide improved air quality by taking into account increasing demand without compromising the mobility of the population.

The impact of transport control measures on emissions is typically measured as the reduction of vehicle emissions these strategies bring about. Many transport models now incorporate technologies for measuring pollutants generated by road traffic, to assist in the evaluation of transport strategies and take into account their respective environmental impacts.

## Coupling traffic and emission models

Generally, due to the difficulty of taking measurements or due to the uncertainties associated with many transport scenarios, vehicle emissions are estimated by combining emission and traffic models. This coupling is usually used to assess the environmental effectiveness of traffic strategies and their potential before implementation (Jie et al., 2013).

Traffic and emission models generally have three levels of spatial representation: macroscopic (traffic flow), mesoscopic (group of vehicles) and microscopic (individual vehicles). In the temporal representation, these models can be classified as static and dynamic. Static models basically assume that traffic status and emissions are stationary during the period analyzed. They consider traffic movements and not the way that they occur. Dynamic models describe traffic flows and therefore represent traffic situations encountered by vehicles during their travel (Cappiello, 2002).

Estimating road transport emissions requires complete information on traffic characteristics such as vehicle fleet composition and traffic conditions. In this context, it is necessary to use appropriate models to accurately estimate traffic and emissions to closely represent the reality. Static emission models are coupled with static traffic models for application to large-scale studies. The static traffic model is one of the most common types of traffic model used to generate inputs, sometimes aggregated in space and time, in order to produce emission models. They are widely used since they can be applied effectively in larger urban areas with low computational effort (Tsanakas et al., 2017).

The study conducted by (Tsanakas et al., 2017) showed that static traffic models cannot reproduce the dynamic phenomena of transport and result in underestimations reaching up to 40% of pollutant emissions. Despite the large number of well-founded methods available and useful for measuring emissions and their concentrations, and evaluating emission mitigating strategies, this approach can mask considerable heterogeneity in the impact of policies on urban areas in which traffic-related emissions are subject to substantial spatial and temporal variability. This fact highlights the need for more detailed measurements of traffic, pollutant emissions and air quality combined with adapted methodologies, and for including them in the policy making process to strengthen policy packages and ensure benefits for air quality (Frecht et al., 2015). Good understanding of traffic dynamics is fundamental to facilitate choosing the most effective study strategy to be adopted for each type of problem being treated. In this context, microscopic simulations can reproduce the effect that any changes will have on traffic, and predict its resulting behaviour. They can also be instrumental in defining the appropriate strategies to be adopted to improve the traffic in question (Schiper et al., 2016). Furthermore, emission models must be sensitive to the effects of traffic dynamics on traffic emission estimations. In this thesis we focus on microscopic models of traffic because it is at the local scale, especially in urban areas, that traffic dynamics have the greatest impact on emission estimations.

## The issues of coupling traffic and emission models and estimation precision

Taking into account the effects of traffic dynamics on networks and more precisely the accuracy needed, leads to significant increases in the volume of data processed and the calculation time required to obtain results. This complexity is necessary when it comes to creating a high resolution description of emissions as they evolve in space and time. It may seem excessive, but it is fair to compare different projects in relation to their global impacts.

Many emission models exist and are fed by representations of the traffic specific to each scale of approach (Can and Leclercq, 2009). Only the most detailed models are really able to take into account the effects of traffic dynamics, e.g., congestion phenomena, and therefore evaluate the environmental impact of different traffic control strategies. The more precise the description of the traffic phenomena in the input of an emission model is, the larger the calculation to quantify the emissions will be, and the finer the spatial and temporal resolution. This calculation of very large quantities of data may seem superfluous when we are only interested in the total quantification of atmospheric emissions in a given area, but is nevertheless necessary to take into account the effects of traffic dynamics and local variations on traffic conditions. The question, however, is whether an effective sampling method would not achieve the same results by keeping a precise description of the phenomena only for a sub-sample. This thesis specifically addresses this issue in order to improve methods of assessing the impacts of road development projects and traffic control strategies.

Errors in the estimation of vehicle emissions may lead to the implementation of traffic management strategies or the organization of transport that are not necessarily the most efficient in terms of air quality. Therefore, errors linked to coupling emission and traffic models must always be related to the gain expected by the implementation of traffic strategies that may be less than 4% (Font and Fuler, 2016). The accuracy of emission estimations obtained from coupled traffic and emission models is affected by two types of errors: (i) errors associated with the reliability of traffic data, and (ii) errors inherent to the modeling of vehicle emissions. These two points are also addressed in this thesis.

Consequently, this thesis focuses on data traffic sampling to significantly reduce the volume of data to be processed while achieving an accurate estimation of the overall results in terms of air pollution. The aim is therefore to define the minimum sample in time and in space as a function of the emission model. For example, rather than making calculations for each part of the network, a set of links and reference time periods will be identified to perform the calculations. The main challenges represented by the objective are related to the integration of spatial and temporal correlations between traffic data and, more generally, to the inclusion of the temporal dimension. The spatial-temporal correlations are linked to waves of congestion and changes in demand that propagate through the network. Thus it is important to define a methodology able to take into account the correlations for the segmentation of the population to define a representative sample. Moreover, traffic dynamics and conditions change on different time horizons (from minutes to a day). It is important to correctly estimate both the emissions and the time during which they occur. This may lead specific sampling differentiated according to the periods of a day.

This thesis comprises five chapters. Chapter 1 presents a review of the existing literature on different traffic and emission models and the different approaches taken to couple these models. Chapter 2 presents a sensitivity analysis of the accuracy of the representation of traffic as a function of information sources and how it affects emission estimations in urban areas. It also includes a descriptive analysis of traffic and emission data obtained from the network used throughout this thesis and their correlations. Chapters 3 and 4 focus on different sampling methods for estimating traffic and emission on various spatial and temporal levels. Finally, chapter 5 presents the analysis of the uncertainties inherent to the emission model and how it propagates in the emission estimations on different spatial-temporal scales.

# 1 Estimating road emissions from traffic models

Estimating atmospheric emissions from road transport requires satisfactory knowledge of traffic data. This estimation is often done by combining traffic models with emission models due to measurement difficulties or uncertainties associated with new transport scenarios (Can and Leclercq, 2009).

Traffic models can forecast the position and kinematics parameters of all vehicles and emission models can estimate the amount of various pollutants emitted by them, while accepting a certain level of uncertainty.

Traffic models are classified into 2 main categories: static and dynamic. Methods and tools are presented for different spatial and temporal scales, input data and results. For example, a static traffic model estimates flow and average speed on a link, while a microscopic dynamic model is able to determine vehicle position, speed and acceleration over all the roads considered. According to many authors, emission models can be divided into categories according to the data needed to calculate emission rates and the different types of pollutants emitted by vehicles. For example, the COPERT emission model can use the results of a dynamic traffic model (vehicle kilometer traveled and average speed) to estimate the emission rates of a large number of pollutants from various emission sources (exhaust, non-exhaust and evaporation).

This chapter presents a bibliographic review of existing traffic and emission models to provide better understand of how they function. First of all, the characteristics of each model, their use and limitations are presented. Next, this chapter describes how these two types of model are coupled according to the use and compatibility of the data in order to estimate road emissions. Finally, the scientific barriers associated with such coupling approaches are presented as the objectives of this thesis.

## 1.1 Description of model

### 1.1.1 Traffic models

Due to its characteristics, traffic is a field where the application of models can play important roles in analysis and for aiding decision. This basically due to two main reasons: (i) the cost and difficulty of carrying out full-scale experiments, not forgetting possible safety implications; (ii) the possibility that the models offered have to be able to predict, test, evaluate and compare several alternatives before they are implemented (or not).

The main objective of traffic models is to predict the effects of users' decisions as a function of a given demand, by trying to reproduce the performance level of the network. According to (Scherr,

2003) to do this it is necessary to:

- obtain estimates of traffic volumes, speeds and delays;
- obtain estimates of aggregate network variables (average speeds, total delay, road emissions, fuel consumption, etc.);
- estimate travel times between zones;
- identify congested links or zones;
- identify the main itineraries between zones;
- analyze zones that use a given itinerary or route choice.

Traffic models can be classified according to the level of detail with which they represent the flow of vehicles. This categorization is carried out by considering the vehicles distinguished and the level of their description: microscopic models represent individual vehicles; mesoscopic models represent vehicles as platoons; macroscopic models represent the vehicles in an even more aggregated way, representing them as a continuous flow. Moreover, depending on the temporal scale, these models can be static or dynamic. Static models focus more on the spatial distribution of the population to calculate average traffic volumes in different zones of the network. According to (Fellendorf and Vortisch, 2000), most static models are based in four characteristics: travel generation, trip distribution, modal split and assignment. All trips are used to build the origin-destination (OD) matrix. The elaboration of an OD matrix involves gathering and crossing sets of socioeconomic data that are adjusted and calibrated with field surveys, in which travelers are interviewed at predetermined locations to identify various attributes of their journey, such as origin, destination, reason, mode of transport used, etc. By combining the OD matrix with other traffic information such as modes of transport and speed flow curves, it is possible to calculate the travel times in the links of the network. These models are highly simplified and they focus only on vehicle flows over spatial scales and do not really consider traffic situations encountered during the journey.

Dynamic models describe the evolution of the state of traffic over time. In general, these models can describe the spatial and temporal evolution of traffic conditions. They can also provide the location and other vehicle parameters by time step and are used to predict, for example, pollutant emissions as a function of space and time. The choice of modeling approach depends on the objective of the study and the constraints in terms of data availability and computation time (Hoogendoorn and Bovy, 2001). Another class among the dynamic models was identified by (Geroliminis and Daganzo, 2007), namely aggregated dynamic models. These models describe the temporal evolution of traffic states (*e.g.* the representation of congestion) but in a more simplified network, thus taking into account spatial aggregation. Basically, they divide the network into reservoir zones based on trip length and also consider a homogeneous distribution of the traffic. Thus it is possible to estimate the average speed and the level of congestion level over time. Essentially, these models are based on the relationship between displacement per unit of time (*i.e.* generation) and the number of vehicles between the different zones in the network (*i.e.* accumulation). This relationship between flow and density is known as MFD (Macroscopic Fundamental Diagram). This type of model is often used for evaluating congestion to reduce vehicle traffic.

According to (Hoogendoorn and Bovy, 2001), traffic models can be classified according to:

- the scale of independent variables (continuous, discrete or semi-discrete);
- the scale of spatial application (networks or links);
- the representation of processes (stochastic or deterministic);
- working process (analytical or simulation);
- level of detail (macroscopic, mesoscopic or microscopic).

The scale of independent variables is considered in static and dynamic traffic models. In static models, it is assumed that the variables are constant over the time period under analysis, while dynamic models allow their variation. The latter can be classified as continuous, in which the traffic state changes continuously over time, or as discrete, where system state changes occur in a discontinuous way at discrete time intervals. Regarding the scale of application, the dimension refers to the spatial study of networks (urban or inter-urban), sections of links or isolated intersections. In the representation process, the mode of behavior of the vehicles in a network can be represented on the basis of two approaches, namely: (i) stochastic models, which reflect the random and probabilistic nature of the variables, and (ii) deterministic models, in which equal behavior is considered for all the actors. Thus it is a process that describes the variables in an analytical way without taking into account randomness. As for the operation process, the models can be defined as analytical, in which all the relations between the system variables are obtained by mathematical equations, or simulation when the variables are obtained using simulations.

Taking last place in the classification of simulation models is the level of detail. The level of traffic representation is traditionally determined by the combination of temporal and spatial scales, as shown in figure 1.1.

Macroscopic models use aggregate variables, such as vehicle flow, average speed, and density, to describe traffic as a flux without distinguishing vehicles individually. Individual vehicle behaviors such as lane changes are generally not explicitly represented (Chakroborty, 2006).

The so-called “four-step models” are the best-known macroscopic and static models. The generation of mobility flows, their spatial distribution, and modal choice with the traffic assignment are the classic approach taken by this type of modeling (Sétra, 2012). Examples of macroscopic and dynamic models are the LWR model proposed by (Lighthill and Whitham, 1955) and (Richards, 1956), the METACOR model (Elloumi et al., 1994) and the CMT model (Cell transmission Model) (Daganzo, 1994). Basically, these traffic models describe the evolution of traffic over time through a set of differential equations derived from the theory of hydrodynamic flow. These models consider real flow dynamics whereas static models are concerned only with vehicle flows.

By contrast, microscopic models use disaggregated variables to represent the movement of individual vehicles within the traffic flow. Vehicle trajectories, driver behavior and their interactions are detailed by their instantaneous position, speed and acceleration. This model depends on the behavior of the drivers, the characteristics of their vehicles and traffic situations (*e.g.* free-flow or congestion), with the most commonly used variables being individual speed, time and distance between vehicles. The basic theory of these models stems from the interaction between vehicles, and considers that the driver reacts concordantly with the following vehicle. Consequently, this provides a more "realistic" representation of the way vehicles circulation in the network; however, it requires a larger quantity of data and more computation time (Hoogendoorn and Bovy, 2001). Currently, there is a wide variety of microsimulation programs such as the PARAMICS model in the United Kingdom (SYSTRA, 2016),

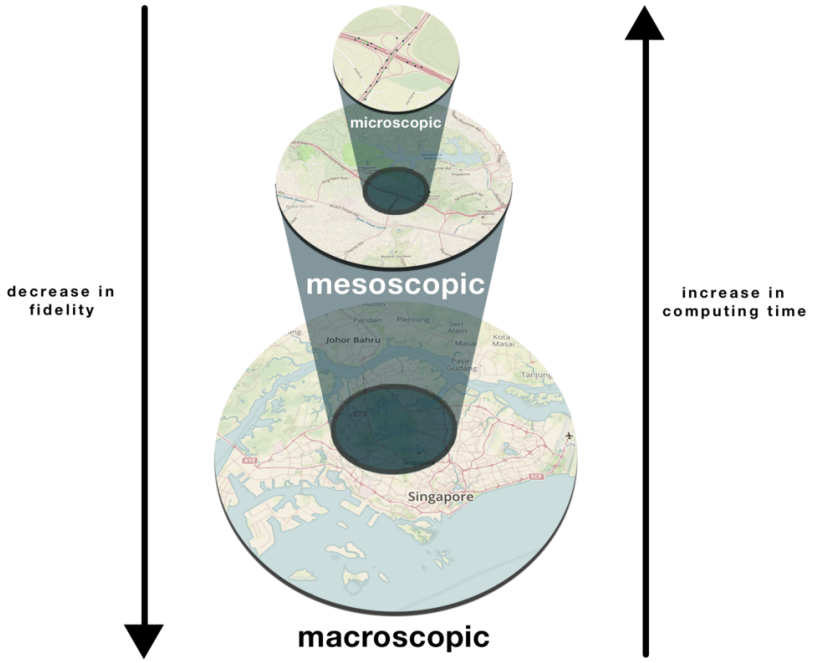



Fig. 1.1 – *Level of detail of the representation of vehicle flows (Zehe, 2015)*

.

the AIMSUN model in Spain (Barcelo et al., 1998), the CORSIM model (United States) (McTrans, 2017) and the VISSIM model in Germany (Fellendorf and Vortisch, 1998).

The mesoscopic models describe traffic flow at a medium level of detail, combining the properties of both microscopic and macroscopic models. Vehicles and driver behavior are not described individually, but rather in more aggregated terms (*e.g.* using probability distribution functions). However, behavioral rules are described at an individual level (Hoogendoorn and Bovy, 2001). There are two main types of mesoscopic models: those in which the vehicles are represented by small groups of vehicles characterized by the same properties of mobility in the network (*e.g.* the same origin and destination), and those in which the dynamics of traffic flow are determined by the simplified dynamics of individual vehicles.

Exhaustive traffic simulators are available that can represent each level of detail but they cannot be dealt with here. Some of them have been reviewed by (Boxil and Yu, 2000) who analyzed their strengths and weakness.

The applications of the macroscopic and mesoscopic models are limited to cases where certain local phenomena do not become preponderant, to define the functioning of the whole system. They have the advantage of describing the individual behavior of the vehicles, without needing to describe their behavior over time (*i.e.* long time periods) and space (*i.e.* large-scale network). More precisely, macro-modeling can be applied to study high-density traffic, but it does not lend itself easily to traffic situations where there is a large variation in behavior between drivers. According to (Burghout, 2004), mesoscopic models are commonly used for traveler information systems. The same author highlights that mesoscopic models are built with a large amount of parameters, which make their calibration and applicability difficult in real time.

One of the main advantages of microscopic models is their ability to observe the displacement of an individual vehicle. Microscopic models are generally appropriate for local studies because, on a larger scale, they require long computation times and calibration is difficult to achieve. Their applications include (i) impact studies of traffic strategies on traffic flows and vehicle kinematics, and (ii) modeling intelligent transport systems (Hoogendoorn and Bovy, 2001).

The evolution of computer capacities means that microscopic models have become more accurate and adaptable to special cases, thus this type of model is used more and more. However, increasing simulation detail hinders the calibration process and increases the probability of modeling errors, so that in the case of unforeseen circumstances, simulation applications generally require a greater expenditure of resources than initially foreseen. Thus, simulation projects must be properly planned so that the resources expended by the modeler are used in the most productive and effective way. (Ortuzar and Willumsen, 2011) studied the various types of errors that can cause incorrect traffic estimations: (i) errors from measurements or incorrect recording of information; (ii) errors from sampling estimations which depend on the number of observations. Some sampling strategies were studied by (Daganzo, 1980); (iii) errors from iterative computation procedures; (iv) errors from simplifications of the process in the models, such as the misrepresentation of a traffic phenomenon; (v) so-called transfer errors which describe cases when a model was developed for a specific case and is applied to another one completely different one; (vi) errors from data aggregation that cannot represent individual behavior. The same author also showed that complex models produced better results compared to the simplest ones because they reduce errors from specifications and data measurements.

In this thesis, we are interested in microscopic dynamic models that provide adequate traffic conditions for more accurate estimations of road emissions. In particular, we are going to use the Symuvia simulation package developed by the LICIT laboratory. It is based on the Lagrangian resolution of the LWR model (Leclercq et al., 2007)) for the car-following law and multiple extensions (lane-changing, multiclass, intersections...) to address all the characteristics of urban traffic. Regarding emission estimations, some traffic models have modules that estimate pollutant emissions. However, these modules present considerable variations both in their theoretical approach and in relation to the emission factors associated with vehicles. Efforts to develop numerical traffic modeling packages rarely contemplate updating vehicle emission parameters. Estimates of vehicle emissions in the emission modules of traffic models should be considered carefully. The following section describes how to consider emission models, and their advantages and disadvantages.

### 1.1.2 Emission models

The objective of emission models is to quantify vehicular emissions based on the mode of operation of the vehicles concerned. Such models have also been used to determine the amount of emissions generated by traffic due to the difficulty of quantifying them in the real world. They are developed using measurement data on emission rates obtained from vehicles. There are several methodologies available for quantifying pollutant emissions. According to (Sturm et al., 1996) the definition of the type of emission model depends very much on the specific need and accuracy required to describe the behavior of road traffic emissions.

Emission models are used for two types of calculation. They can be used to predict absolute values of pollution (inventories), such as identifying streets that exceed air quality standards, but for this type of analysis a high degree of precision regarding emission factors is required. Emission models are also used for impact assessments, such as the comparison of different traffic strategies to reduce

emissions. In this type of analysis, the precision of the emission factors may not be a very important factor.

The main pollutants traditionally modeled by vehicle emission models are: carbon monoxide ($CO$), volatile organic compounds ($VOCs$), nitrogen oxides ($NO_x$), carbon dioxide ($CO_2$) and particulate matter ($PM_{2.5}$ and $PM_{10}$).

Nitrogen oxides ($NO_x$) are not direct products of combustion. However, their production takes place in an environment created by combustion. It is due to the chemical reaction between the nitrogen present in the atmospheric air and the gases of elevated temperature formed by combustion. The $NO_x$ emitted are composed of nitric oxide ($NO$) and nitrogen dioxide ($NO_2$), the latter is significantly smaller in quantity than the first. When fuel consumption is low, a small amount of $NO_x$ is emitted. The emission of $NO_x$ in diesel engines is higher than that of gasoline engines. This is due to the combustion characteristics of diesel engines, which have higher temperatures and pressures (De Nevers, 2000). $NO_x$ is one of the precursors of ozone formation. In addition, it reacts with ammonia and other components to form nitric acid, which can cause respiratory problems.

Carbon monoxide ($CO$) is formed during the combustion process. Its concentration is directly linked to the air/fuel equivalence rate (De Nevers, 2000). It is produced in the case of incomplete fuel combustion. This occurs when there is insufficient oxygen ($O_2$) to burn all the carbon ($C$) contained in the fuel. Diesel engines operate with excess air, and consequently produce insignificant amounts of $CO$ compared to that produced by the gasoline engines (De Nevers, 2000). $CO$ is an invisible and odorless, but very toxic, pollutant. In the human body it reacts with the hemoglobin present in the blood, causing a reduction of the $O_2$ transported to the cells. Prolonged exposure to $CO$ can cause dizziness, headache and even choking, depending on its concentration. In addition, high concentrations can cause heart and respiratory problems in children and the elderly.

Volatile organic compounds ($VOCs$) are formed by incomplete combustion or evaporation of the fuel (particularly during refueling). Due to the higher volatility of the fuel in gasoline engines, they emit higher proportions of $VOCs$ than diesel engines.

Particulate matter ($PM$) consists of solid or liquid substances (unburnt carbon fuel particles) that can be collected by the filtration of exhaust gases. In gasoline motors the emission of PM is insignificant compared to diesel engines, where much of this material is generated. PM is characterized by its size (coarse, fine particles, and inhalable particles).

Carbon dioxide ($CO_2$), which is the main greenhouse gas emitted by vehicles, and whose emission is related to the energy efficiency of the vehicles (fuel consumption). It is also the main product of complete fuel combustion.

The emissions produced by a vehicle have a high degree of variability. Several factors affect the level of a vehicle's emissions. Emission variability manifests itself in two different ways, that between vehicles and that in the emissions generated by the same vehicle. The variability of the emissions from one vehicle to another presents a high order of magnitude and results from technological factors, and vehicle wear and maintenance. The variability of the emissions of a single vehicle is dependent on environmental, operational and, in some cases, maintenance conditions (Barth et al., 2000). The latter is presented in more detail in the next section.

### The main factors influencing vehicle emissions

This section describes the mechanisms and impacts caused by all chemical compounds emitted by motor vehicles. The relative amount of each chemical is a function of a various factors, including

engine technology, fuel type and driving behavior.

*Technological factors*

The technological factors are grouped into three categories according to (Barth et al., 2000); (i) emission control equipment; (ii) fuels; (iii) motor type. Emission control technologies have been incorporated over the last 30 years, including the recirculation of exhaust gases to reduce $NO_x$ formation in the engine, the adoption of the catalytic converters for the treatment of exhaust gases, the replacement of the carburetors by electronic fuel injection and computer controlled air-fuel mixing and ignition timing. Generally speaking, fuels such as diesel tend to generate a higher amount of emissions, and the reduction of this type of fuel by more sustainable fuels is essential for increasing air quality. In general, emissions are very sensitive to these technologies regardless of the age of the vehicle. Vehicle characteristics such as engine volume, power and weight also affect emissions rates.

*Wear and maintenance factors*

As a vehicle ages and its cumulative mileage increases, its emissions tend to increased. This phenomenon is a function of both the natural degradation of the emission controls of vehicles with good conservation status, resulting in moderate increases in emissions over time, and of the malfunction or failure of emission controls resulting in a considerable increase in emissions ($CO$ and $VOCs$). According to (Wenzel et al., 2000), the distribution of the emissions of a large number of vehicles is very distorted. Most vehicles have relatively low emissions, while a relatively small number of vehicles with functional problems have extremely high emissions.

*Operational and environmental factors*

Average speed is the operational variable used most to describe the level of emissions. Average speed is a combination of speed and acceleration of each type of highway. This is because the engine operating regime is significantly different for each type of highway given the same average speed. For example, an average speed of 50 km/h on an urban road gives the idea of free flow, while on a highway this same average speed indicates a congested flow with frequent accelerations and decelerations, and consequently a higher level of emissions.

Accelerations play a major role in emissions. Observed acceleration levels are strongly correlated with the aggressiveness of the driver. Of the total emissions generated by a vehicle on a trip, mots are composed of small episodes with high emissions. These small episodes occur in acceleration events (Rouphail et al., 2000). Levels of emissions from stationary vehicles are very low. The influence of acceleration events on the formation of emissions was also studied by (Rakha and Ding, 2003).

The effects of acceleration on emissions are most noticeable on urban roads. Traffic is smoother on rural roads. An aggressive driver emits up to 8 times more emissions than a moderate driver in vehicles equipped with a catalytic converter (Vlieger et al., 2010).

The study conducted by (Leblanc et al., 2005) reported increases on $CO$ emissions for vehicles exceeding a speed of 90 km/h in all acceleration intervals, as well as when the speed was lower than 90 km/h but at which acceleration exceeded 5.3 km/h/s. Remote sensing studies in Houston have linked vehicle exhaust emissions to the level of instantaneous activity, noting that emission rates were a function of speed and acceleration (Yu, 1998).

(Rakha and Ding, 2003) quantified the impact of stoppages on vehicle emission levels. Maintaining the same average speed, $VOCs$ emissions have a significant impact due to a stop (100% for an average

speed of 80 km/h). The lower the average speed, the lower the impact of stopping on emissions.

According to (Ahn, 2002), emission rates are influenced by the physical characteristics of highways. Facilities such as signal intersections, inclinations of roads, toll gates and traffic interlacing sections can increase the level of emissions as a function of the engine due to accelerations. The slopes of highways also affect the emissions. On an inclined road, a vehicle needs more engine power in order to maintain vehicle speed. Also, road surface conditions (*i.e.* irregularity) influence the amount of emissions emitted.

*Other factors*

Characteristics such as the travel time, ambient temperature, humidity and altitude influence emissions levels. According to (Ahn, 2002), in Denmark, although the proportion of driving time with vehicles with cold engines corresponds to only 9% of the total time of all trips, cold starts contribute 60% of $CO$ and $VOCs$ emissions. This phenomenon is minimized in countries with a tropical climate, where the average temperature is higher than in the countries of the northern hemisphere. Very low temperatures influence emissions at vehicle start-up and cause catalyst cooling even at very short stops (Wenzel et al., 2000). Air humidity can affect the level of $NO_x$ emissions. Furthermore, altitude has a strong influence on the emissions of diesel vehicles.

Among these main factors, we focus on the kinematic variables of the vehicle as they are used as input traffic data for emission models and play an important role in estimating these environmental externalities.

### Model development and measurement methods

The driving cycle is designed to represent a typical driving pattern in a region and is widely used in emission studies. Several driving cycles have been developed by institutions in various countries. According to (Barlow et al., 2009), these cycles vary considerably between each other, since they seek to reproduce specific driving behaviors influenced by local traffic conditions and the characteristics of the route traveled. These test cycles to which vehicles are subjected are primarily intended for certification purposes, making it possible to compare all vehicles under similar conditions. The duration of a driving cycles varies from one-two minutes to thirty minutes or more, usually given from second to second. Thus, in order to reduce testing costs per vehicle, driving cycle development efforts are challenged to simulate real vehicle behavior for specific categories of vehicles, roads and speeds (Smit et al., 2010).

A chassis or engine dynamometer is used for tests of both engines and vehicles. It simulates the resistive power imposed on the wheels of a vehicle, and is capable of applying speeds and controlled loads to measure an engine or vehicle in terms of torque, power, movement, etc. Vehicle-based emission measurements result in a specific unit of grams per kilometer (g/km) and established emission factors that can be used to estimate pollutant emissions according to the speed.

The development of vehicular emission models is based on measurements of unit emission factors in vehicles from the driving cycle tests. Emissions can be estimated by linking these unit factors to the vehicle operating mode (Sétra, 2012).

$$E_i = O \times EF_i \tag{1.1}$$

where:

- $E_i$ is the emission of a given component $i$, generally expressed in mass (g/km);
- $O$ is the vehicle's operating mode (*e.g.* distance traveled);
- $EF_i$ is the emission factor of a given component $i$ per kilometer;

Emission models can be classified as dynamic or static according to the time scale used, and as macroscopic or microscopic according to the spatial scale. The choice of modeling approach depends on the objective and the constraints of the study. According to (Sturm et al., 1996), the method adopted to obtain these emissions factors varies according to the model approach described below:

*Static models*

Models based on average speed are those used most commonly and take into account vehicle dynamics through the concept of average speed. They work based on specific emission factors for each type of vehicle/engine technology, degradation factor and an average traffic situation. Generally, they form the basis for calculating air quality on a local scale. They are characteristically employed for large scales such as cities.

They take shape according to vehicle usage statistics, such as annual mileage, road types, and so on, and calculate average emissions, including the effects of cold starts, evaporations, etc. They are also used for regional and national emission inventories. These models cannot be used to generate instantaneous emission estimations since they determine emissions by time as a function of the average speed of a cycle. The applications of this type of model include large-scale analyses and cases where the average speed adequately characterizes the flow of traffic (*e.g.* continuous flow on highways).

In addition, in some situations the same average speed may correspond to different operating conditions (Smit et al., 2010). However, the most relevant argument is that the driving cycle, used for the development of models based on average speed, presents operating characteristics (average speed and accelerations) that differ considerably fromreal world operating conditions. To offset these deficiencies, the latest versions of emission models include 11 or more types of driving cycles, according to route classification and service level. A few examples of models are MOBILE (EPA, 2004), EMFAC (CARB, 2002) and COPERT (Gkatzoflias et al., 2012).

*Dynamic models*

In the dynamic approach, emissions are measured continuously in chassis dynamometer tests and stored at specific time intervals (usually every second). The operating conditions of the vehicle at each time interval, usually the speed and acceleration value, are observed simultaneously with the emissions. These instantaneous measurements allow instantaneous and modal analysis of emissions based on instantaneous kinematics variables, such as speed and acceleration, or more aggregate modal variables such as time spent in acceleration mode, cruise mode or stopped mode.

Instantaneous models have two sorts of classification. Regarding the description of traffic, they can be: (i) microscopic, when they employ instantaneous kinematics variables (speed and acceleration second to second); and (ii) mesoscopic, when kinematics variables are aggregated (*e.g.* speed and number of stops, speed and mean acceleration, time spent in each operation mode).

Instantaneous models can be classified according to the methodology used for their development. Emission maps take the form of a matrix in which one dimension represents the speed variations and the other accelerations or power. The calculations are performed from the visualization of the

matrices, where each speed and acceleration corresponds to a specific amount of emissions. Regression-based models are generally linear regressions that employ acceleration functions and specific speeds as explanatory variables. Physical models represent the physical and chemical phenomena that generate vehicle emissions. These models are composed of modules that simulate each step of the process. Some examples of dynamic models are PHEM and HBEFA (Hausberger et al., 2009), MOVES (EPA, 2010).

A major advantage of dynamic models is that they take into account congestion phenomena as they require the input of actual driving situations or specific congestion variables to determine emissions. Theoretically, they can model the effects of congestion on emissions. However, a large quantity of data is needed to provide the information necessary for sufficient accuracy (Sétra, 2012).

### Limitations of emission models and uncertainties

The accuracy of emission model outputs depends on the uncertainties on the internal parameters of the emission model, such as emissions factors and the input data. The main causes of these uncertainties can be ambient conditions, vehicle fleet composition, vehicle mileage, traffic data and emission factors.

The uncertainties on the emission factors stem from statistical errors in their calculation, imperfections in sampling and analytical methods considered as measurement errors, differences between the source test and the case study, average time for measurement, reference data, missing data, the statistical estimation used to complete missing data, and others.

For example, the COPERT model calculates the emission factors that describe the amount of pollutants produced by one vehicle per km. Basically, the model considers two types of emission factor: hot emission factors and cold emission factors. This differentiation stems from the fact that the amount of pollutants produced by a vehicle depends directly on the engine temperature. Hot emission factors correspond to the amount of emissions produced under stabilized engine temperature conditions, while cold emission factors refer to the pollutants generated during the period when the engine has not yet reached the appropriate temperature.

After calculating the emission factors, it is possible to calculate the hot and cold emissions that a vehicle produces in a year from the quantity emitted in a km, in hot or cold conditions, multiplied by the number of km traveled per year, and by identifying the mileage in hot or cold engine temperature conditions.

Identifying the number of km traveled under hot or cold conditions is complex. To achieve more satisfactory results, COPERT calculates and uses two correction factors: a mileage degradation factor, which considers the age of vehicles and assumes that older vehicles emit more pollutants than new vehicles, and a real fuel factor, which considers the effects of improved fuels used on older vehicles, in which case such vehicles produce less pollution than older vehicles with ordinary fuels.

The uncertainties associated with this model were studied by (Kouridis et al., 2010). The same author studied fifty-one uncertainties on inputs in different case studies that can be a source of error and the most important sources were identified as meteorological and temperature parameters, vehicle fleet composition, mileage, vehicle speed, average trip length and fuel properties. Also, COPERT is widely used in air quality modeling studies because it can cover the major emission processes and most of the pollutants.

## 1.2 Coupling traffic and emission models

The previous sections presented the characteristics of each type of traffic modeling and emissions individually, and they are shown in figure 1.2. The association of traffic models to provide information to support emission estimates can lead to benefits since the events contributing to emission levels can be evaluated at a detailed level.

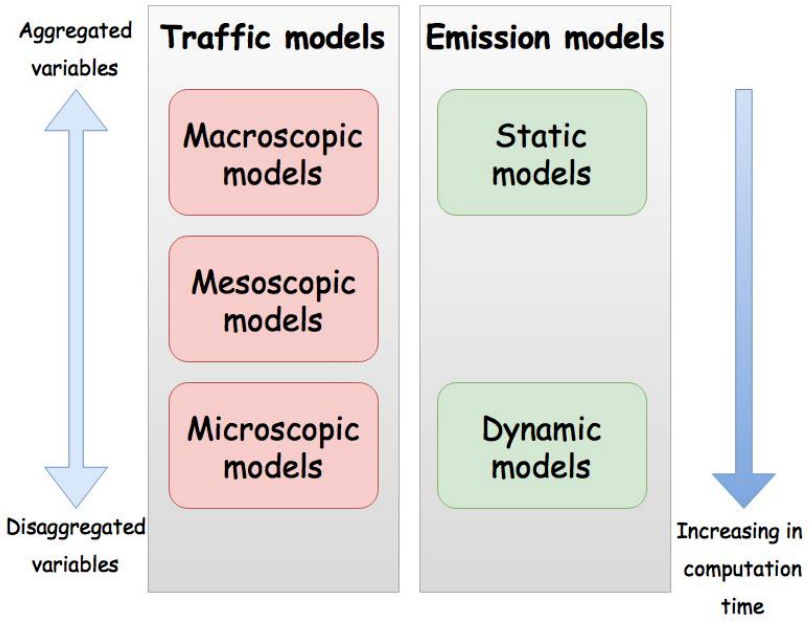



Fig. 1.2 – *Representation of traffic and emission models (Rocha, 2013)*

.

The use of microscopic traffic simulation has grown due to technological and computational advances. In addition, traffic model data is increasingly used for estimating vehicle emissions. In addition, many of these traffic models have their own built-in emission models, such as DRACULA (ITS, 1993) and INTEGRATION (Van Aerde, 1999).

Static emission models can use data from macroscopic or mesoscopic traffic models. Vehicle-kilometers traveled are determined from surveys and used as inputs for emission models. Although this coupling cannot provide a precise and disaggregated output, this is commonly used for transport planning purposes because of its relative simplicity (Ahn, 2002). The study by (Zhang, 2011) estimated $CO_2$ emissions from vehicle mileage traveled (VMT) data coupled with vehicle emission factors. This methodology is used to evaluate mobility strategies to reduce these emissions.

Static emission models can also use data from microscopic traffic models. Different approaches can be used to fuel the emission model with vehicle speeds and accelerations. It is possible, for example, to apply a spatial distribution of speed and/or acceleration based on driving cycles or statistical distributions (Burghout, 2004).

The major disadvantage of macroscopic traffic models is that they do not replicate the traffic peculiarities responsible for the largest variations in emission levels such as high accelerations and decelerations due to congested traffic conditions (Ahn, 2002). As a result, finer approaches are recommended to better assess the effect of traffic variability on the emission estimate. In this context, microscopic models of traffic are coupled to emission models.

In emission modeling, the input data are derived from microscopic traffic models. This type of finer coupling is generally used to test the environmental impact of traffic strategies at the local level

(Ahn, 2002), different configurations for multiple occupancy vehicles (Barth et al., 2000), and different urban traffic control strategies to evaluate emission reduction strategies (Yu, 1998).

The input data of emission models characterizing traffic behavior range from describing the change in speed over time (every second) for each vehicle (instantaneous emission models) and an average speed characterizing a set of vehicles on a more or less wide spatial-temporal scale (aggregate emission models). These traffic data can be obtained by fixed or mobile measurements (*i.e.* counting sensors) or by dynamic traffic simulation. The use of traditional on-road sensors for collecting real-time data is necessary but not sufficient because of they are expensive to implement. It is also a disadvantage that such technologies, for obvious reasons, only provide local information. Methods are therefore necessary to expand this local information to larger spatial scales but they currently suffer from the following limitations: (i) the relationship between missing data and estimation accuracy cannot be easily determined, and (ii) calculations for large areas are computationally expensive. Dynamic traffic simulation makes it possible to have sufficient data sets to test different levels of coupling between traffic and emission models at various spatial and temporal scales.

In this case, a large volume of data must be processed and evaluated to estimate the corresponding pollutant emissions. It is on this last point, the processing and exploitation of traffic data, that the thesis will focus. In other words, the solution proposed consists in estimating the total emissions from the emissions associated with a smaller sample. From this assumption, the work described in the thesis will consist in defining a method of sampling the data representing the traffic so as to reduce the volume of data to be processed while maintaining the reliability of the overall results in terms of air pollution. Similarly, rather than performing an emission calculation for each link of a network, it will be necessary to identify a set of links and reference time periods for carrying out the calculations on the global level. This method will also open new perspectives for the implementation of network-wide monitoring systems designed to efficiently estimate pollutant emissions.

## 1.3 General objectives of the thesis work

Recent research on the environmental impact of transport modeling has used data from traffic models. The bibliographic review of this chapter presented the different traffic and emission models available to provide better insight on the stakes involved in this approach. Afterwards, the relationships between these models with different degrees of detail of the variables were presented. The choice of the coupling approach chosen depends directly on the objective of the study.

It was also concluded that assessing pollutant emissions requires satisfactory knowledge of traffic-related data. The key issue at stake is to accurately describe traffic dynamics and more especially congestion periods. Classical methods for assessing traffic-related pollutant emissions are based on an aggregated description of vehicle behavior. Static emission models such as COPERT require only mean speed and travel production (total distance travelled by a vehicle over a given period of time) to estimate the related emissions. Indeed, the total emissions are calculated as the product of the vehicle-kilometer traveled and the unitary emission factors (based on vehicle technology or vehicle fleet) that depend on mean speed.

However, when the traffic description stems from a static approach, the consequences of traffic congestion are often poorly estimated. Congestion periods induce considerably lower speeds. It is therefore important to obtain accurate estimates for both traffic mean speed and vehicle-kilometer traveled when calculating emissions. Microscopic dynamic traffic simulators can provide good esti-

mates of these two variables directly but at the cost of expensive computation time.

The question, however, is whether an effective sampling method would not achieve the same results by maintaining a precise description of the phenomena, but only for a sub-sample. This thesis specifically addresses this issue in order to improve methods of assessing the impact of road development projects and traffic control strategies.

In this thesis, the objective is to define the optimal selection method to estimate emissions. In relation to this objective, the main deadlocks are linked to taking account of the spatial-temporal correlations between traffic data and, more generally, taking account of the temporal dimension. Spatial-temporal correlations are related to congestion waves and changes in demand that propagate in the network. Thus, it is important to define a method capable of taking into account these correlations in view to segmenting the population and define a representative sample. These studies and their methodology are structured as follows:

**Chapter 2**

Presentation of the microscopic model of traffic and the emission model used throughout the thesis. In addition, an investigation is made into the sensitivity of pollutant estimations to errors, considering traffic input data from different sources at the urban scale;

**Chapter 3**

Presents the datasets used in this thesis and a complete analysis of the traffic and emissions estimated using a linear statistical sampling method;

**Chapter 4**

Presents a comparison between naive and complex statistical sampling methods used to estimate network emissions on different temporal scales;

**Chapter 5**

Presents a study of the influence of the spatial and temporal aggregation of traffic input data used to estimate pollutant emissions.

# 2

# Descriptive analysis of a network transportation - Paris case

This chapter is the only part of this study focused on simulated data. A neighborhood of Paris, part of the $6^{th}$ district, was used as the basis for our study. The network was built with the help of a dynamic microscopic simulator. Geo-referenced maps provided by the national institute of geographic and forest information (IGN)were also used as part of the project ISpace & Time funded by the ANR. Geo-referenced maps have a fine description of the geometry, typology and physical characteristics for the delimited neighborhood. The network is composed of 234 links, 93 crossroads, 19 entries, 21 exits, 4 car parkings and 27 traffic lights. All the links have directions, bus lane, traffic light times and allowed turning movements in crossroads. The traffic settings were defined during the ISpace & Time project. The detailed description of how simulations were conducted, the hypothesis considered, the variables, the emission calculations and other important considerations that were taken into account are explained through this chapter.

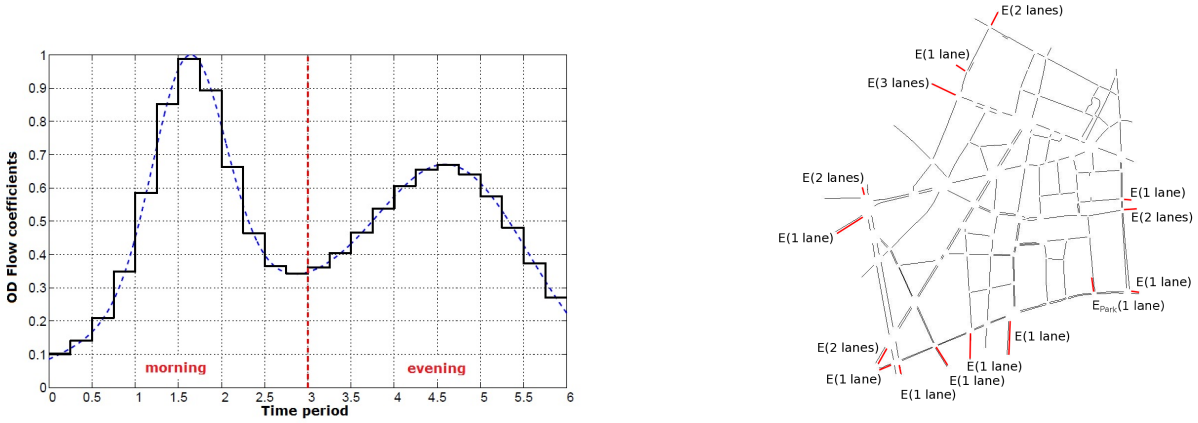
## 2.1 Traffic simulations

### 2.1.1 Simulation environment

The dynamic traffic simulator Symuvia was used to define the traffic settings which represent, in a most realistic way, the traffic conditions on the neighborhood. Three main settings were taken into account: the temporal evolution of the demand, the origin-destination conditions and the rooting scheme within the network. In order to avoid long calculation time to simulate 24 hours of traffic, the 6 most relevant hours for typical daily traffic are considered. The temporal evolution of the demand is represented by two peak hours of the traffic: the morning and the evening. The first one corresponds to the intense demand distributed in short periods while the evening peak experiments a lower demand peak but during a longer time.

As can be seen in figure 2.1, the demand varies every 15 minutes for each network entry. The x-axis represents the three pertinent hours in the morning and the evening traffic and, the y-axis represents the coefficients used to calculate the demand for each entry. The demand is the capacity defined by the product of the coefficients and the maximum level of demand. This last one is calculated using the macroscopic fundamental diagram taking into account the network structure (e.g. mainly axis with high traffic demand) and the maximal level of demand in this network can reach a demand value equal to 0,7183.

Figure 2.2 shows the maximum level of demand considered for the entries that have a high level of demand, their IDs (i.e. blue IDs means entries and red IDs means exits) and also the traffic light




(a) The temporal evolution of the demand

(b) Network entries

Fig. 2.1 – *OD flow coefficient for each time period and the entries where the demand values are applied.(Villegas et al., 2013)*

positions.

Once the demand is calculated, the flow distribution from each entry to each exit needs to be determined. To introduce variability, the flow distribution varies around 15% for each time period considered, namely origin-destination matrix (OD matrix), and additionally, the flow direction changes completely between morning and evening. The flow direction is represented in figure 2.3, the arrow sizes stand for the flow magnitude in the direction indicated.

The last setting is called assignment matrix or route choice. For each OD couple, there are many possibilities of routs to reach a destination from an origin. The percentage of vehicles that use a determined routs for each OD couple needs to be settled. To this end, the multinomial logit model was implemented in Symuvia. This model affects the flow for a given OD couple as a function of their cost as shown in equation 2.1. The cost of route choice is represented by the travel time and the $\theta$ parameter, both can be different for each OD couple.

$$p_i = \frac{e^{\theta C_i}}{\sum_{k \in L_{mn}} e^{-\theta C_k}} \tag{2.1}$$

where,

- $C_i$ is the itinerary cost;
- $p_i$ is the flow proportion of the OD couple $(m, n)$ for the itinerary $i$;
- $L_{mn}$ is every possibility of itinerary from origin $m$ to destination $n$;
- $\theta$ parameter is the shape of the itinerary function distributions. For our study we set $\theta$ equal to 0,05.

Incorporating all these parameters in the microscopic simulator, it is possible to obtain an accurate definition of the traffic physics in the network for each time step, for any setting definitions and for any network.

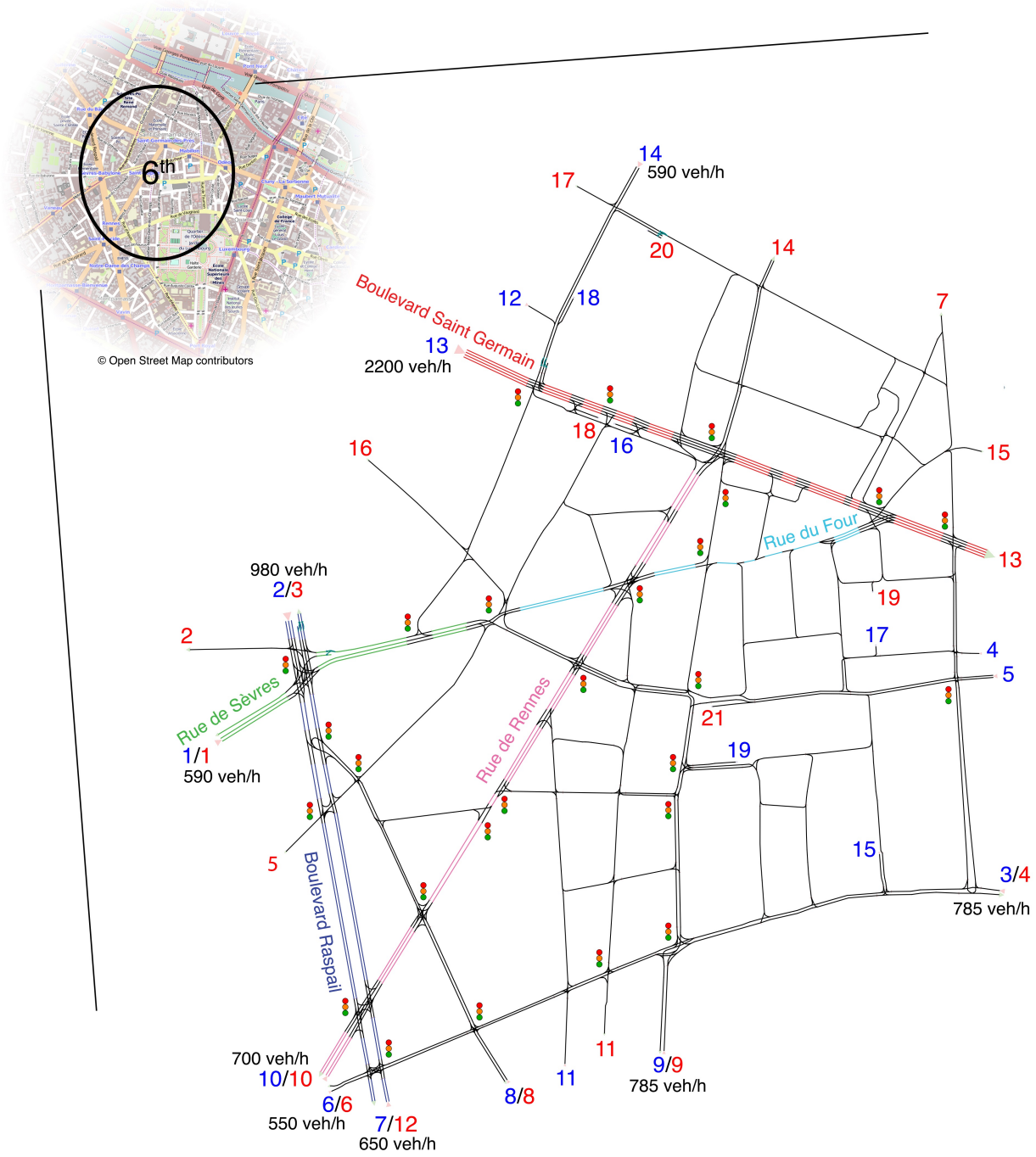



Fig. 2.2 – *The network profile made in Symuvia with the traffic light positions. The major axes of the network are indicated with colors. The demand values are shown in each entry, and they correspond to the traffic peak hours of the demand values.(Villegas et al., 2013)*

### 2.1.2 Data Generation

In order to simulate the traffic in the proposed network, only passenger cars were modeled. At link level, two kinds of information are stored: one that provides the spatial information (namely over the whole link) and the second that provides local information namely information observed in the middle of each link. The minimum local information that may be obtained by inductive sensors in real cases. Both have traffic time aggregation of 15 minutes, which means that the 6 hours of traffic

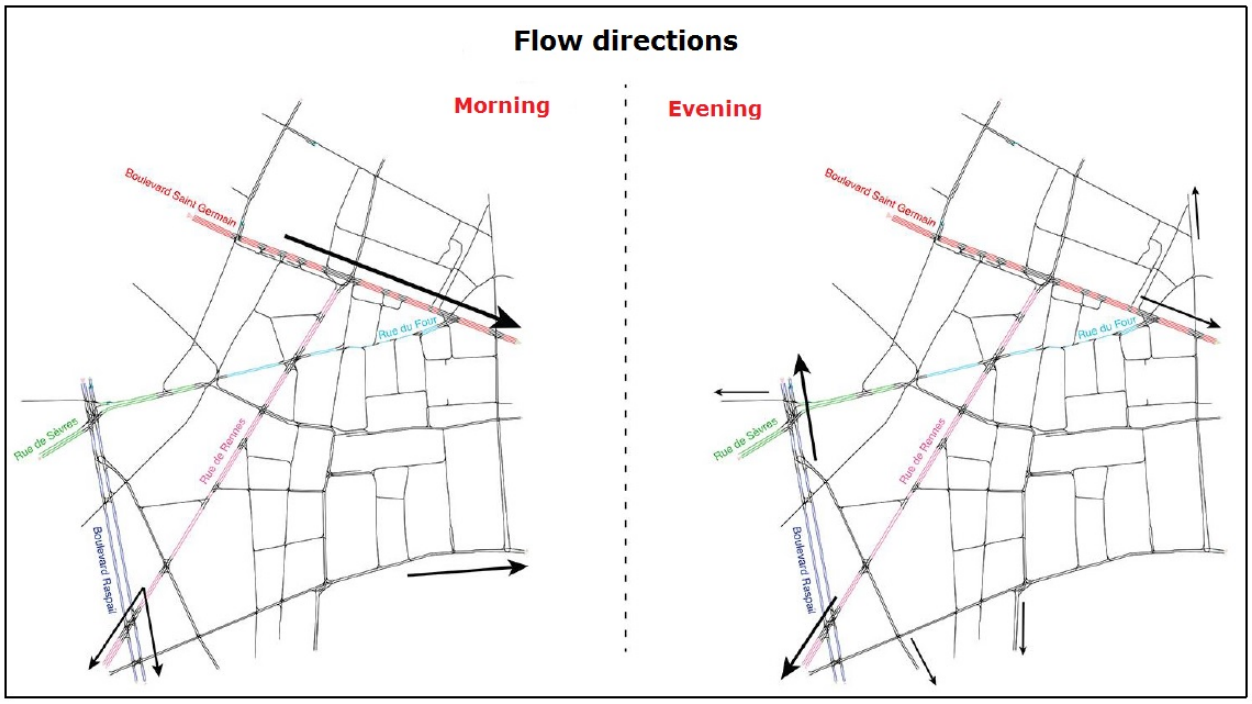



Fig. 2.3 – *Intensity of flow directions considered in the morning and the evening time. (Villegas et al., 2013)*

simulation were divided into 24 periods of 15 minutes. For each period and link of the network, the traffic information was recovered. All traffic data used in this work came from the traffic simulator Symuvia using the microscopic model as explained in chapter 1.

Spatial information provides a complete description of the local traffic as: the travel time, the traveled distance by cars on the link and their spatial mean speeds. Local information dispenses vehicle flows and mean speeds in the middle of each link. This sensor is an example of traffic information recovered type used by policymakers to evaluate their strategies to regulate networks. Indeed the traffic simulation gives two different sources of traffic data. It is important to note that spatial traffic information at link level, in reality, is not available, but it is possible only through simulations. A sensibility analysis between both will be discussed later on this chapter.

In order to be statistically representative, a great number of observations (i.e. simulations) and a varied traffic state in space and time are needed. To this end, the number of simulations were set at 400, resulting that the traffic simulator was launched 400 times to represent more than a year of traffic data. The difference between them is the demand value for each entry and time period. The average value of demand was defined in the project ISpace & Time (Villegas et al., 2013). These values were then varied randomly in $\pm$ 50% for the 400 simulations, resulting in a demand matrix with values in space (entries) and time (periods) respecting the shape of distribution defined in the project as shown in 2.1. Figure 2.4 shows the demand values by time. Each simulation is represented by a color line. The Boulevard Saint Germain is the axis with major flow of the network. The origin-destination and assignment matrix remain unchanged between all simulations.

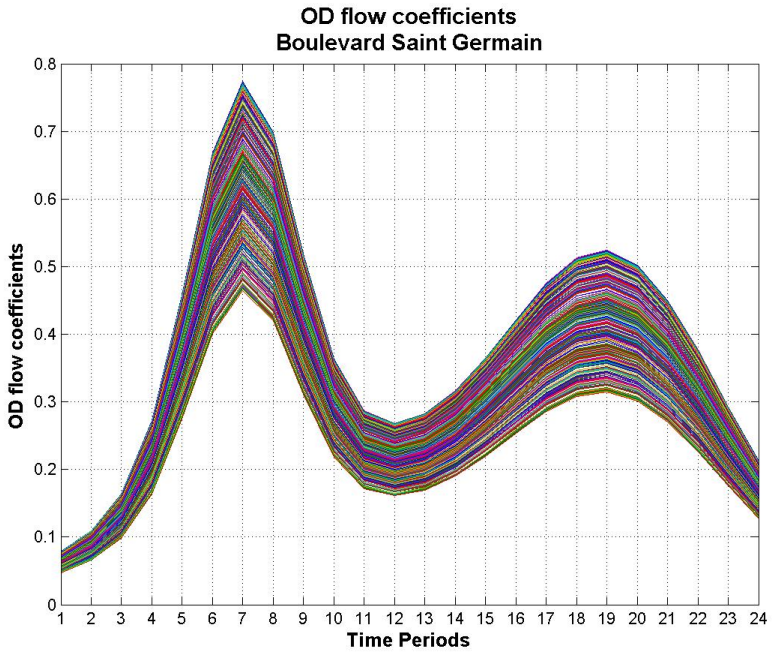



Fig. 2.4 – *OD flow coefficients for each period of time in Boulevard Saint Germain.*

## 2.2 Traffic and pollutant emission variables

In this section the traffic and emission variables will be presented and how they will be considered. The traffic variables are the traveled distances and the spatial mean speeds. For emissions, the $CO_2$ (carbon dioxide) and $NO_x$ (nitrogen oxides) will be calculated using traffic variables. These variables will be evaluated in two different ways: one from spatial information (i.e. spatial sensors) and the other one from local information (i.e. punctual sensors). Their definitions will be explained in the next lines.

Once all simulations were launched on Symuvia, the traffic data for each sensor was recovered with the help of Matlab software. Four traffic variables were recovered from the spatial sensor, from each time period in every link: the travel times,the total traveled distances, the spatial mean speeds and two definitions of link lengths. The total travel time and total traveled distance stand for the total time spent by vehicles, respectively the total traveled distances between beginning and end of each link during the time period settled. The spatial mean speeds are calculated directly by the simulator for every link and time period as the ratio between the total traveled distance and the total travel time.

The punctual sensors give only vehicle flow and mean speeds at their location (i.e. in the middle of each link in every link of the network). In order to calculate emissions using the COPERT IV model, two traffic information are required: the traveled distances and mean speeds. Using the local sensor as a traffic information source, the traveled distances can be calculated thanks to the vehicle flows. In Symuvia traffic simulator, the network geometry is composed of two distinct elements: the links and crossroads. Both allow to determine the traveled distances by cars inside the network. To this end, two link length definitions will be used to calculate the traveled distance. The latter is obtained by multiplying the link length by the vehicle flow, resulting in two possibilities of traveled distances by time period. These two definitions of link lengths are called: static and dynamic length. The first called static length ( identified as $l_{bc}$) considers the length between the beginning and the

end of the link plus the distance between the exit of the link and barycenter of the crossroad. The second length (identified as $l_{dyn}$) possibility is called dynamic, has the same definition as the static one but instead of considering the crossroad geometry, it takes into account the distances traveled inside the crossroad. The latter definition allows to know the real distance traveled by vehicles inside the crossroad, according to the flow, instead of estimating them using geometric measurements.

As described above, the local information will be used as input in the emission model to calculate the local emissions. Two pollutants will be considered, the $CO_2$ (carbon dioxide) that has most impact on the greenhouse effect and the $NO_x$ (nitrogen oxides) which impacts the public health. The emission assessment is done according to the choice of parameter settings such as fleet composition, type of emissions and speed-dependent emissions. The 2015 French fleet composition was chosen and a study will focus on hot emissions established by this fleet. To calculate the amount of pollutant, the speed-dependent curve will be used. The latter provides emission factors for each average speed bigger than 10 km/h. Thus, two hypotheses were retained: (i) the emissions from stopped vehicles will not be taken into account; (ii) for average speeds between 1 km/h and 9 km/h, the emissions will be calculated using the emission factor equal to 10 km/h. The equation defining the assessed emissions is shown below:

$$e_p = d_i \times EF_p(\bar{v_i}) \tag{2.2}$$

where:

- $e_p$ is the pollutant emission in g/km;
- $d_i$ is the traveled distance in the spatial element $i$ (link);
- $EF_p$ is the pollutant emission factor associated to a passenger cars fleet determined by average speed in $i$;
- $\bar{v_i}$ is the average speed in $i$.

Figure 2.5 represents how the traffic variables derived from both types of sensors were used to evaluate the respective emissions.

All those variables were used to study how the traffic information source can influence the emission estimations. This analysis is described in the next section.

## 2.3 Influence of the variable definitions

### 2.3.1 Comparison between spatial and punctual sensors

Two possibilities to recover traffic data from simulations are from the spatial and punctual sensors. The variables provided by spatial sensor give a precise data information that represents what happened on the network, for example the exact traveled distances or spatial mean speeds in a given link. In practice, this type of information can only be obtained through traffic simulations and cannot be observed in reality. These values were used as a reference to compare with local traffic information. Later, the differences between both, in terms of variable values, will be analyzed to understand how these differences spread when emissions are calculated.

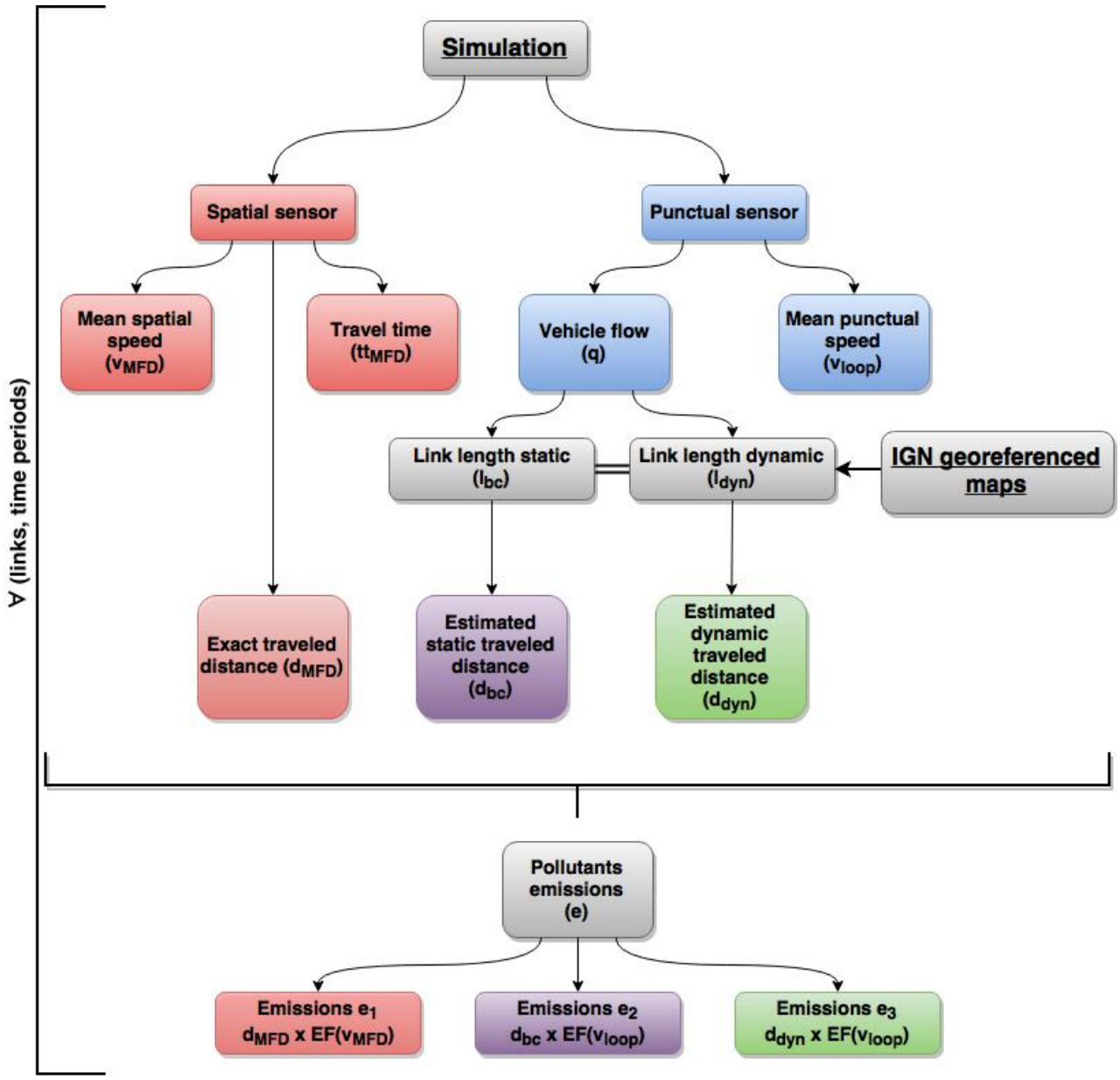
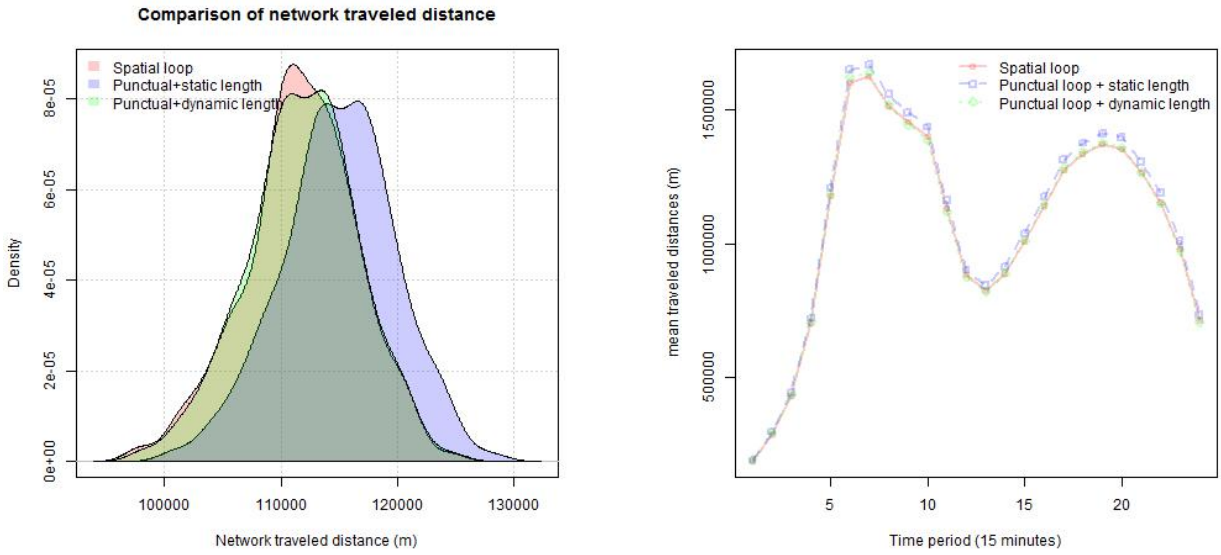



Fig. 2.5 – *The traffic variables recovered from each link and time period of the network and their respective emission calculations.*

### Traveled distance

The traveled distances from spatial sensors are used as a reference to compare with the total traveled distances derived from punctual sensors. This last one recovers, for each time period, only mean speed and vehicle flow for each point fixed in the middle of each link. To obtain the total traveled distance used as an input in the emission model, two hypotheses are explored: (i) the extended link is evaluated as the Symuvia-link length plus the length to the Symuvia-crossroad barycenter (namely static length - $l_{bc}$); (ii) the extended link length is derived from the distance traveled by cars on the link plus the distance traveled inside the crossroad (namely dynamic traveled distance - $l_{dyn}$). Figure 2.6 compares the impact of these two hypotheses on the total distance traveled in comparison to the reference values.

Both figures in 2.6 make a comparison of the network traveled distance. The values represented are the sum of all the distances traveled by cars considering the 400 simulations (all links and time periods




(a) Daily network traveled distance densities comparison.

(b) Network mean traveled distances by time period.

Fig. 2.6 – *Network traveled distance comparison.*

gathered). The total traveled distances evaluated thanks to the spatial sensors are the reference values, "punctual + static length" corresponds to the total traveled distances considering the static link length ($l_{bc}$) and the "Loop + dynamic length" are the traveled distances considering the dynamic length ($l_{dyn}$). The last one is the calculation method that better corresponds to the reference values with only 1% of the average error (over 400 simulated values). Considering the periods of congestion, the same difference is only 0,2% on average. It is interesting to observe that the traveled distances calculated using the static length have almost the same distribution as the traveled distances calculated using the dynamic length. In fact, this is not surprising because the total traveled distances are the product between the number of vehicles that pass at the sensor in a link and the estimation of this link length. Considering that, it is assumed that all vehicles passed through the sensor traveled throughout the entire link. Consequently, this traveled distance will be a little overestimated, especially in congested periods where the difference can reach 3% on average, as shown in figure 2.6 (*b*). Thus, it considers that all vehicles passed in front of the sensor run through the totality of the geometric link length and sometimes it is not the case. The differences between them are small at network level (i.e. space and time gathered). This distinction can also be seen in figure 2.7 which shows the distribution values of each hypothesis and the relative mean error of each calculation method in comparison to the reference values.

Within a perspective of policymakers and considering the low errors of traveled distances (under 3,5% in average over 400 simulations), the method using the static length allows, in an easy way, to determine the traveled distance of a link or network directly using the data collected by the sensors and geo-referenced maps without having to use dynamic traffic simulations to this purpose.

### Mean Speed

The second traffic variable that needs to be analyzed is the mean speed. The network represents an urban area which experiments low mean speeds and they vary over 15 minutes between 1 km/h and 50 km/h locally. Considering the spatial sensor, the network mean speed can be calculated in any

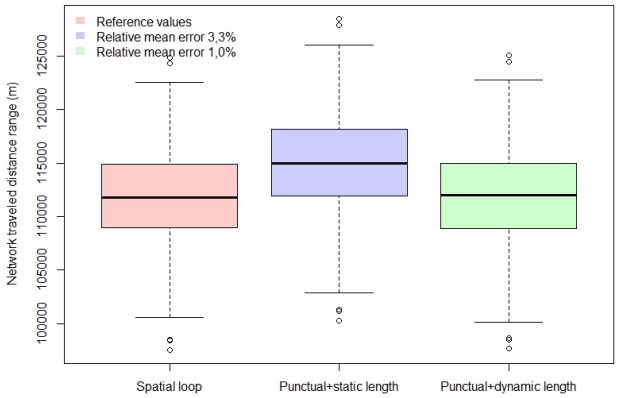



Fig. 2.7 – *The error distributions of daily network traveled distance and their relative mean errors.*

temporal scale. The equation that describes the spatial mean speed is shown in 2.3.

$$\bar{V_{spatial}} = \frac{\sum d^i_{spatial}}{\sum \frac{d^i_{spatial}}{\bar{v^i_{spatial}}}} \tag{2.3}$$

where:

- $\bar{V_{spatial}}$ is the network mean speed from spatial sensor;
- $d^i_{spatial}$ is the traveled distance in the link $i$;
- $\bar{v^i_{spatial}}$ is the mean speed in the link $i$;

The punctual sensors have two possibilities of traveled distances as explained before. The mean speeds were calculated using both estimations of traveled distances.

$$\bar{V_{bc}} = \frac{\sum d^i_{bc}}{\sum \frac{d^i_{bc}}{\bar{v^i_{sensor}}}} \tag{2.4}$$

and,

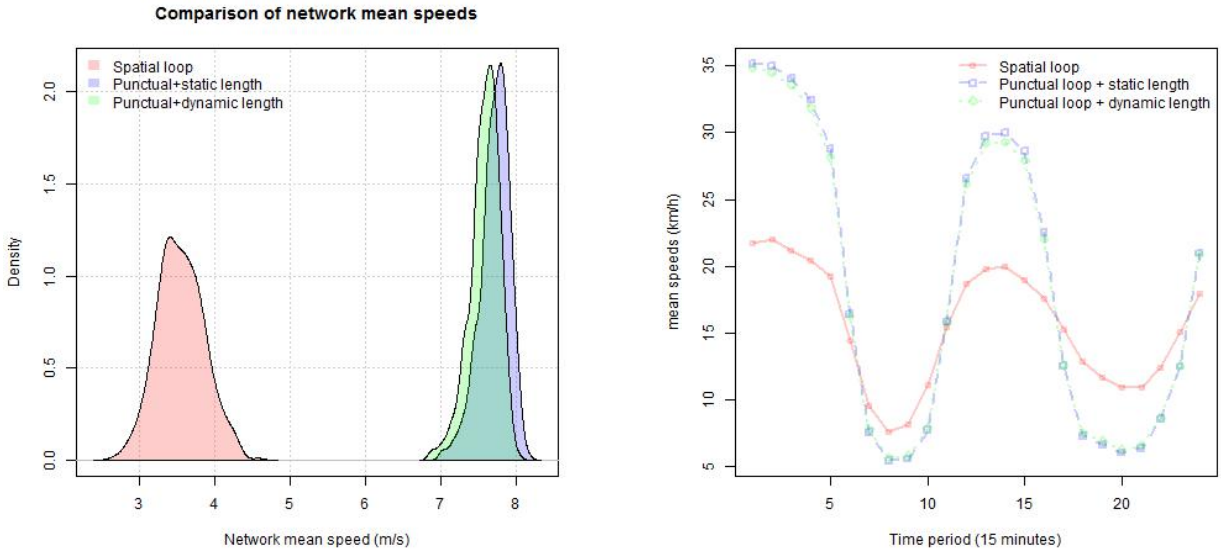
$$\bar{V_{dyn}} = \frac{\sum d^i_{dyn}}{\sum \frac{d^i_{dyn}}{\bar{v^i_{sensor}}}} \tag{2.5}$$

where:

- $\bar{V_{bc}}$ is the network mean speed from the punctual sensor calculated with traveled distance which considers the static link length;
- $\bar{V_{dyn}}$ is the network mean speed from the punctual sensor calculated with traveled distance which considers the dynamic link length;
- $d^i_{bc}$ is the traveled distance calculated with the static link length of the link $i$;
- $d^i_{dyn}$ is the traveled distance calculated with the dynamic link length of the link $i$;

- $\bar{v}^i_{sensor}$ is the mean speed in the link $i$;

The comparison of network mean speeds over all time periods, which were calculated thanks to the spatial sensors in 2.3 and punctual sensors in 2.4 and 2.5 are shown in figure 2.8.




(a) Density comparison of the network mean speeds.

(b) Network mean speed by time period.

Fig. 2.8 – *Network mean speed comparisons.*

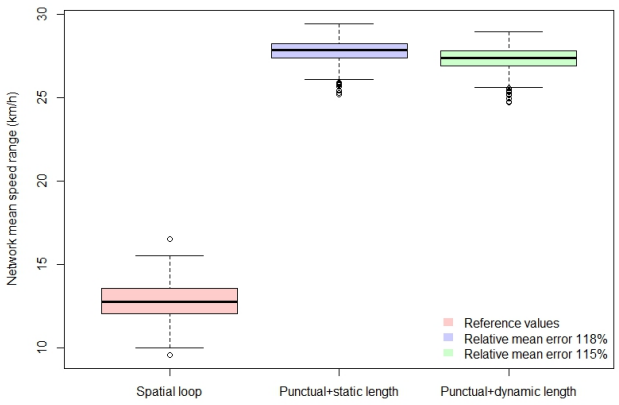


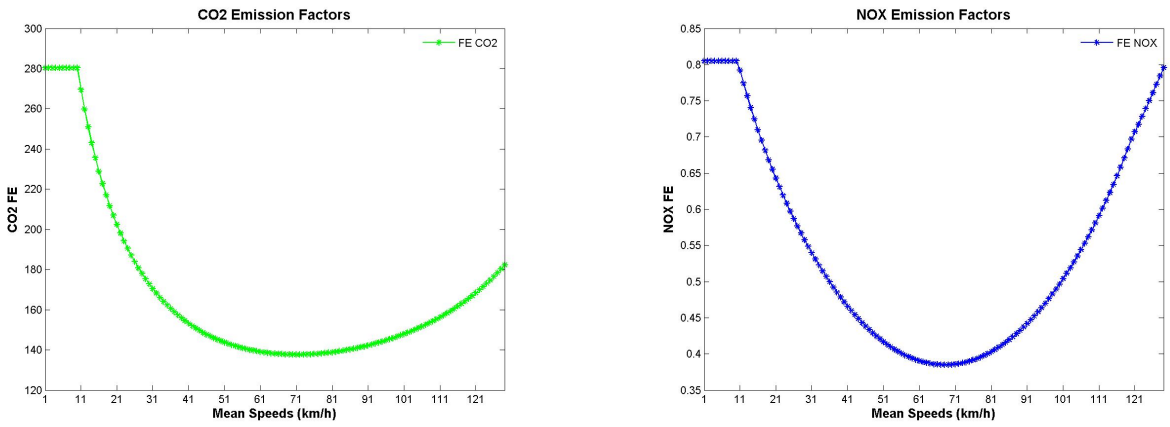

Fig. 2.9 – *The error distribution comparison of daily network mean speed.*

Speeds are the ratio between distances and times and, considering the low difference between static traveled distance (i.e. using geometric link length) and the dynamic one (i.e. using dynamic link length), both give almost the same results. The mean speeds from punctual sensors are both overestimated and reached great relative mean errors, at about 115% on average, as shown in figure 2.9, this is because the punctual sensor location is in the middle of link far from intersection. The range of mean speeds at network scale is very different when spatial and punctual sensors are compared. The range of mean speeds for a time period varies between 5 and less than 35 km/h. These low speeds are totally normal when an urban area is represented. Furthermore, these low speeds have an importance when the emissions are calculated, because the low ones have higher emission factors.

The great differences between the mean speeds from spatial and punctual sensors are more evident in the periods of free flow (spatial sensors better integrate the traffic lights influence) and they can reach 14 km/h of difference between both. This fact is explained in how the mean speed is considered at link level. For punctual sensors, the mean speed considered to the whole link length is measured from a point in the middle of each link. As most links have small length, the vehicles running through the sensor are still accelerating. Unlike the punctual sensors, the spatial sensors calculate the mean speed considering the full-length of each link (spatial approach) and not a point. These considerations explain the differences between mean speeds from both sensors. Moreover, the spatial mean speed can only be obtained thanks to simulations, but it shows that we have to be aware that experimental campaigns will induce a biased estimation of mean speeds.

### 2.3.2 Pollutant Emissions

The emission factors expressed by g/km (i.e. the mass of pollutant emitted by cars per unit of distance) used in COPERT are continuous and often non-linear and they are also specific for different vehicle types, fuel types, hot or cold emissions, etc. The fleet considered is the 2015 French traffic composition taking into account only passenger cars, as they certainly contribute the largest share of road traffic emissions (Smit et al., 2010). The pollutant emissions considered in this study are $CO_2$ and $NO_x$, which means their relative emission-speed curves are used to calculate the amount of pollutant emissions. These curves represent only hot exhaust emissions for passenger cars present in the fleet mix chosen for the study. The emission factors are attributed to mean speeds between 10 km/h and 129 km/h for our case, representative of an urban area. It is common have lower mean speeds mainly when a large number of traffic lights are available and roads are congested. Considering the latter, emission factors below 10 km/h will be the same as that equal to 10km/h. The effects of stop-and-go and cold emission were not taken into account in this study. The emission factor curves considered for this study are represented in figure 2.10. The emission were quantified using these emissions factors




(a) $CO_2$ emission factors

(b) $NO_x$ emission factors

Fig. 2.10 – *Emission factors determined by COPERT IV methodology for passenger cars considering 2015 fleet composition for hot exhaust emissions*

The factor-speed curves from both the pollutants studied illustrate several key ideas. Very low average speeds generally represent stop-and-go driving patterns, and vehicles which do not travel far. Conversely, vehicles traveling at much higher speeds demand very high engine loads that require more fuel and thus lead to high emission factors (Smit et al., 2010). As a result, these curves have a parabolic

shape, with high emission factors at both ends and low ones at moderate speeds between 60 and 80 km/h.

The concept of emission-speed curve can serve as the basis of linking emissions to vehicle activity. In fact, a large family of curves can be established for different roadway types, different levels of congestion, and even for a specific vehicle fleet composition in an urban area. These emission-speed curves can be easily employed with traffic measurements or traffic simulations on a roadway to estimate traffic-related emissions for a specific location. These curves from COPERT were developed through regression on emission data points (*i.e.* the average emission factor for each pollutant over each driving cycle).

To study the influence of traffic variables on emissions, a comparison was made between the amount of pollutants emissions estimated with both traffic information.

**Local approach**

This approach seems to be the most accurate, when the traffic data at the link level is available. The definition of daily network emissions is the pollutant sum on each link and time period. The emissions were calculated using the total traveled distance and mean speed recovered by both sensors for each 15 minutes time range. Then, the emissions were summed, for all links and time periods per daily traffic (i.e. simulations).

This method was used with spatial and punctual information, knowing that the last one has two options of calculated traveled distances, accordingly two possibilities of emission values for each pollutant. For all the studies about emissions, the results from spatial calculation were our reference values, because they used the finest description of traffic and represent the exact values on each simulation.

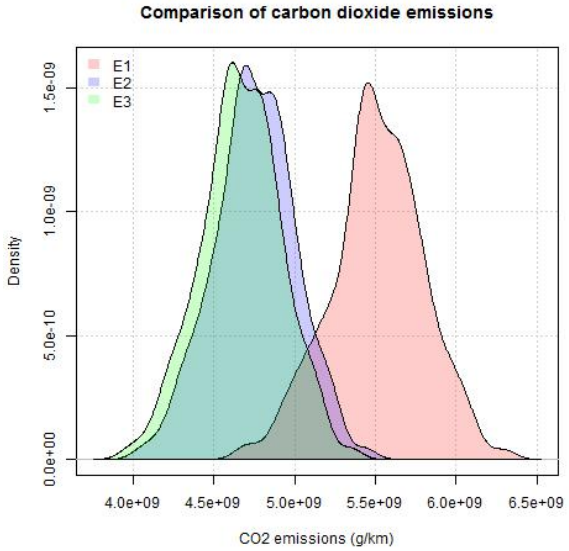
Figure 2.11 compares the pollutant emissions from both sensors: $(a)$ and $(b)$ correspond to carbon dioxide network emissions; and $(c)$ and $(d)$ to $NO_x$ network emissions. As can be seen, the pollutant emissions calculated using traffic data from punctual sensors show lower values than the calculated ones with spatial sensors. These lower amounts of emissions are due to the fact that punctual sensors consider much higher speeds than spatial ones at the link level as explained in 2.3.1 section. These considerations explain the differences between mean speeds from both sensors and consequently the emission values and these differences are most evident in congested state.

As shown in figure 2.8, the network mean speeds from punctual sensors are between 25 km/h and 30 km/h instead of 5 km/h and 18 km/h from spatial sensors, consequently high-speed values tend to have lower a coefficient of emissions, and these are shown in figure 2.10.

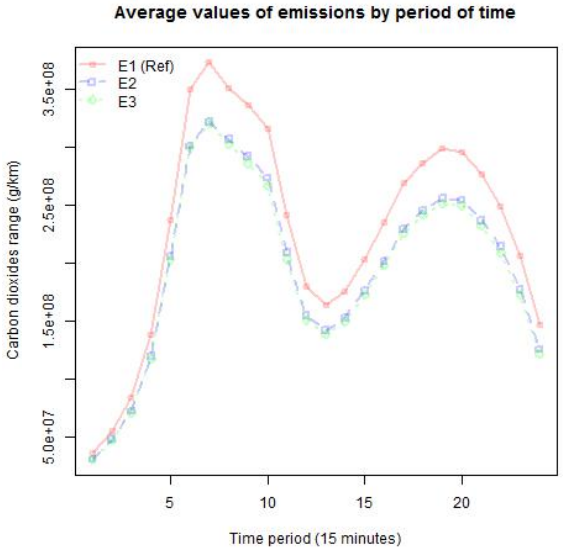
The quantity of pollutant emissions is calculated by the product of traveled distance and the corresponding emission factor for given pollutants determined by mean speed. The difference between the three calculated traveled distances is very small (figure 2.7) but the mean speed comparison shows different speeds from both sensors and that ends to underestimate around 14% the network emissions using the traffic data from punctual sensors ($(a)$ and $(b)$ in figure 2.11). The last consideration can be observed in figure 2.12. So, we will use spatial mean speed for all emission calculations in the study.

**Influence of emission factors**

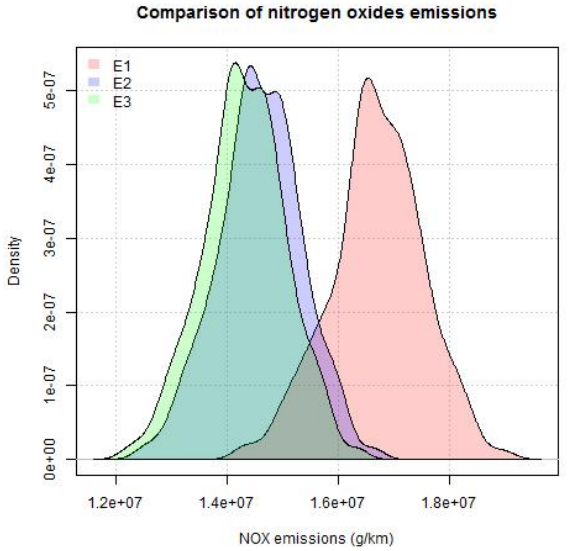
Emissions-speed curves from COPERT IV do not consider emission factors from mean speeds between 1 and 9 km/h, as explained before. Using a curve fitting and approximation functions available in MatLab software, the emission-speed curve for $CO_2$ and $NO_x$ pollutant emissions were fitted to



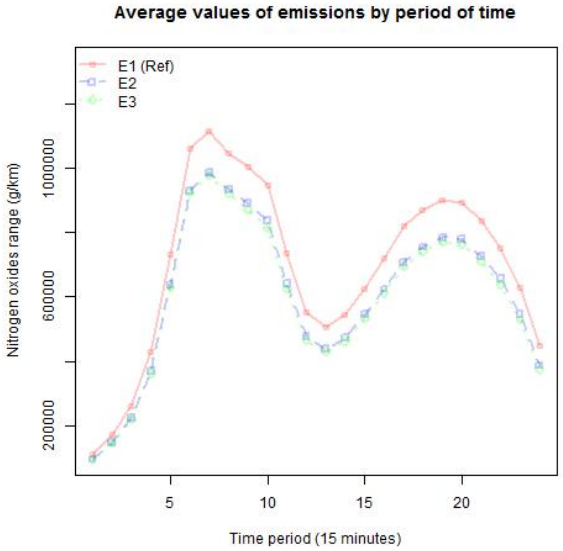

(a) Daily network $CO_2$ emissions densities.




(b) Mean network $CO_2$ emissions by time period.




(c) Daily network $NO_x$ emissions densities.



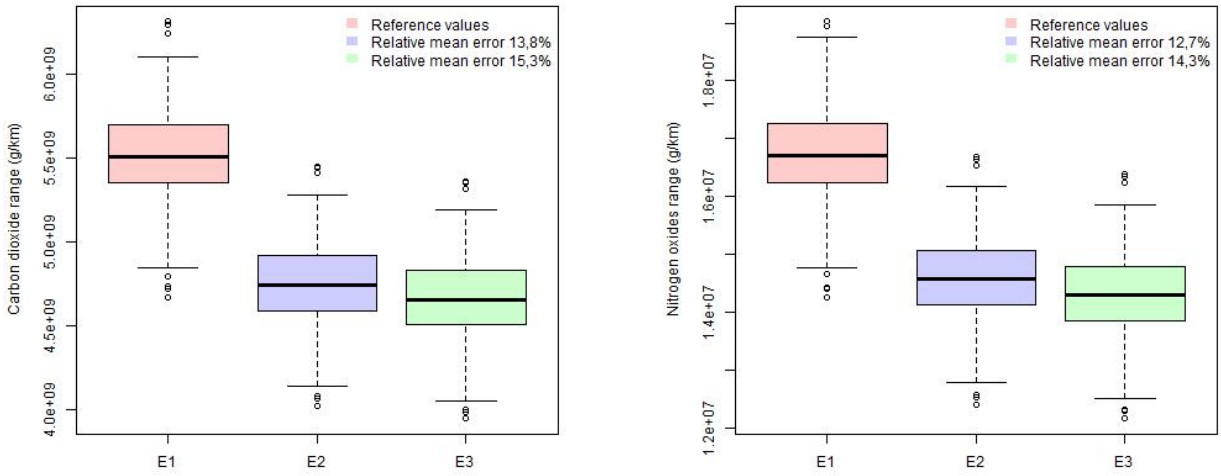

(d) Mean network $NO_x$ emissions by time period.

Fig. 2.11 – *Pollutant emission comparison*

determine the coefficient values for the mean speeds between this specified range. Figure 2.13 shows the curve fitting for each pollutant.

Figure 2.10 represents the assumption that defines for all speeds lower than 10 km/h the emission-factor corresponding to 10 km/h. The fitted curves (hypothesis 2) were used to calculate emissions from the network and have been compared with the first hypothesis. Emissions were calculated for each link and time period separately and then all the values were aggregated to study their distribution, resulting in a total emission quantity in the network for each simulation. The effect of each hypothesis in emissions quantification is shown in figure 2.14.

Figures 2.14 ($a$) and ($b$) show the distribution densities for both pollutants. As can be observed, hypothesis 2 has emission values higher, about 3,1% for $CO_2$ and 2,3% for $NO_x$ emissions, than hypothesis 1. These differences came from periods of congestion on the network where speeds are

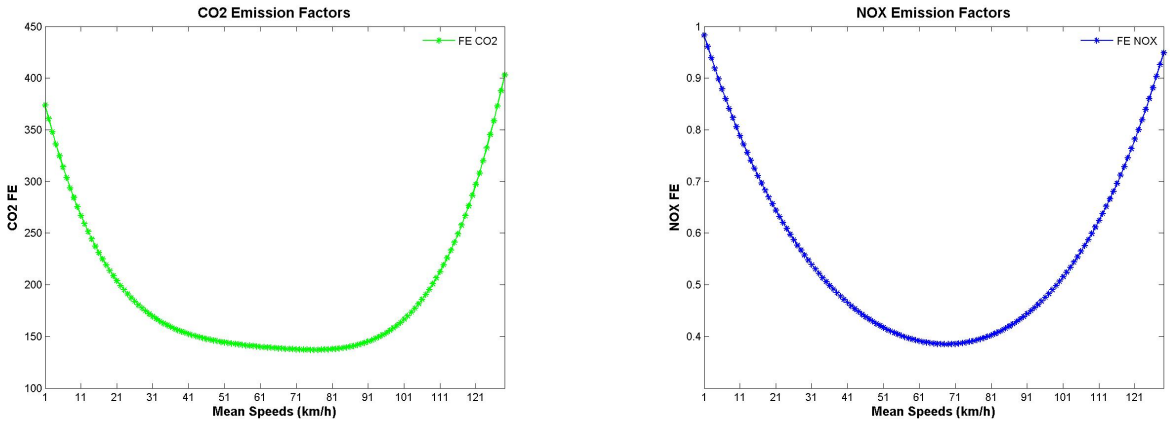



(a) Daily network $CO_2$ emission error distributions.

(b) Daily network $NO_x$ emission error distributions.

Fig. 2.12 – *Comparison of network emissions determined by COPERT IV methodology. E1 represent emissions from spatial sensors (reference values); E2 are emissions calculated using traveled distances determined by static length and; E3 were calculated using dynamic link length to determine traveled distance.*

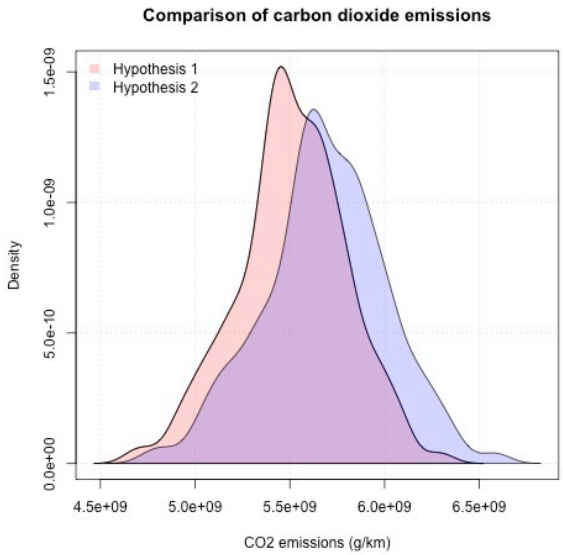
between 1 and 9 km/h, which induces an increase in the emissions on the network. The differences are shown in 2.14 ($c$) and ($d$) which show the average value of network emissions by time period. Analyzing these average values for each pollutant, the percentage of difference between both hypotheses can attempt from $+0,35\%$ in free flow state to $+7,30\%$ in congested period for $CO_2$ emissions. The same happens with the pollutant $NO_x$, this percentage of difference varies between $+0,30\%$ and $+5,14\%$. All those conclusions are presented in figure 2.15. The effect of hypothesis 2 will be more obvious for great congestion and also for lower network mean speeds. In Paris, an urban area, most of the links have low mean speeds, mainly in congested states. Considering that, it is normal to face differences that can reach the disparity of 7% at the network level considering both hypotheses. To continue the



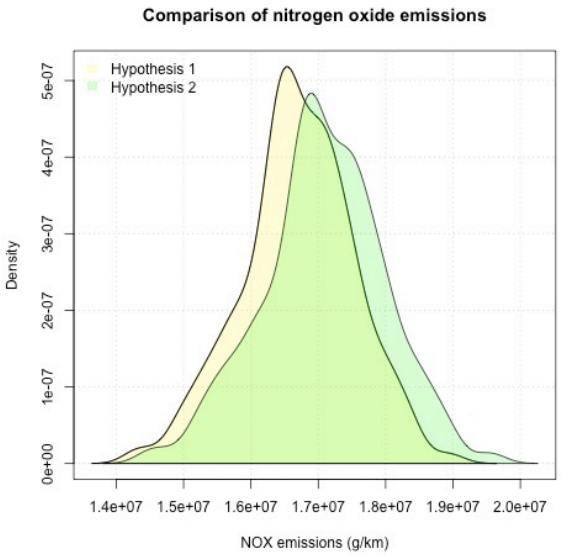

(a) $CO_2$ fitted emission-speed curve.

(b) $NO_x$ fitted emission-speed curve.

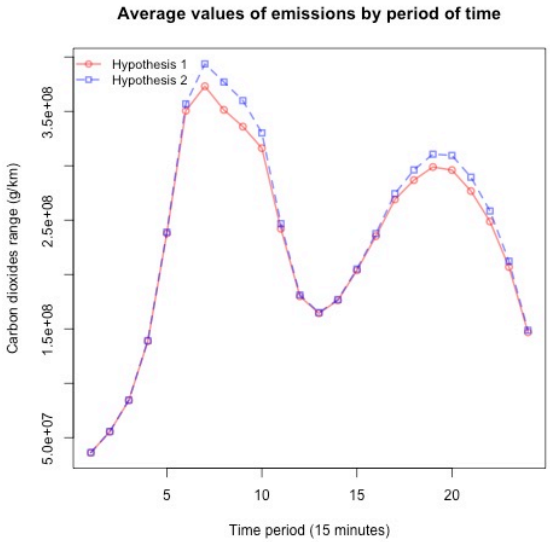
Fig. 2.13 – *Fitted emission-speeds curves.*



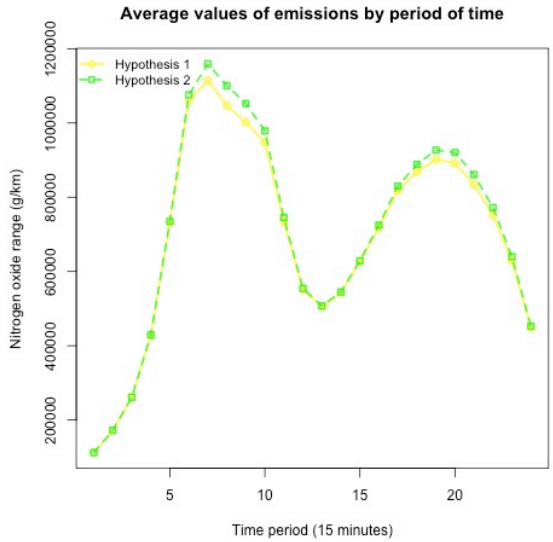

(a) $CO_2$ emission densities.




(b) $NO_x$ emission densities.




(c) Average network $CO_2$ emissions values by time period.




(d) Average network $NO_x$ emissions values by time period.

Fig. 2.14 – *Hypothesis comparison.*

study, hypothesis 1, which uses the factor emission value from the mean speed equal 10 km/h in all average speed under 10km/h, was retained and used as a basis to apply selection methods.

### 2.3.3 Conclusions about the variable definitions

Two sources of traffic information were analyzed: spatial and local information. The first one represents a modeling source and the second represents the traffic data obtained by experimental devices (loop detectors). Both provide the necessary traffic data to estimate emissions. Traffic data from punctual sensors tends to overestimate traffic variables in comparison to spatial ones. Daily network traveled distance is overestimated in average by 2% while the bias on mean speed can reach more than 100%. These gaps lead to an underestimation of daily emissions around 14% at network level in both studied

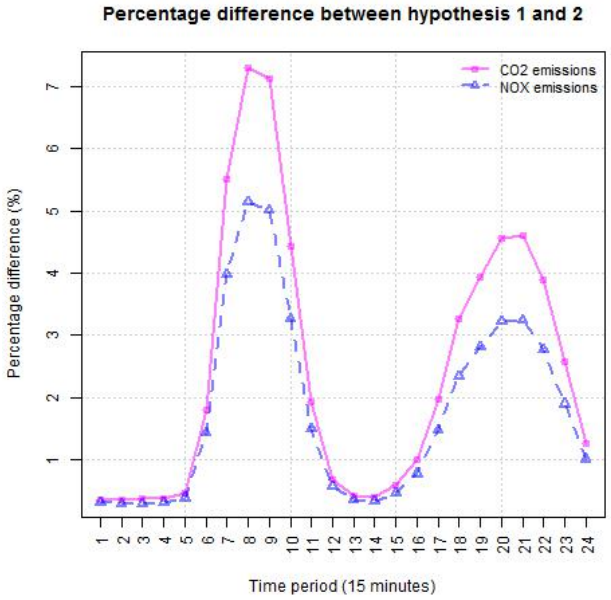
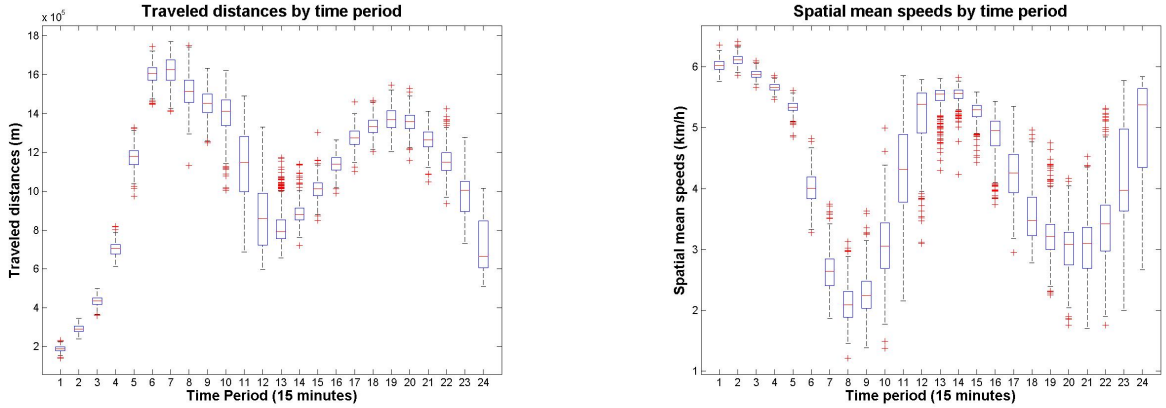



Fig. 2.15 – *Percentage difference between the hypothesis.*

cases. For free flow periods, the disparity between both traffic information is about 1% compared to congest ones which can reach 14% of difference. Also was studied how the emissions factors can influence the amount of emission at network level assigning values for mean speeds lower than 10 km/h. To assess the emissions with more accuracy and to obtain a selection using the accurate values, after having compared all the variables and their impacts emission estimations, traffic data from spatial sensors using local approach will be used as a basis to apply the selection methods.

### 2.3.4 Further spatial variable analysis

In order to produce traffic data of the 6$^{\text{th}}$ district of Paris, 400 microscopic simulations were obtained as explained in section 2.1.2. The distributions of the traffic variables and pollutant emissions are represented through time periods in figures 2.16 and 2.17.




(a) Network traveled distances through time periods.

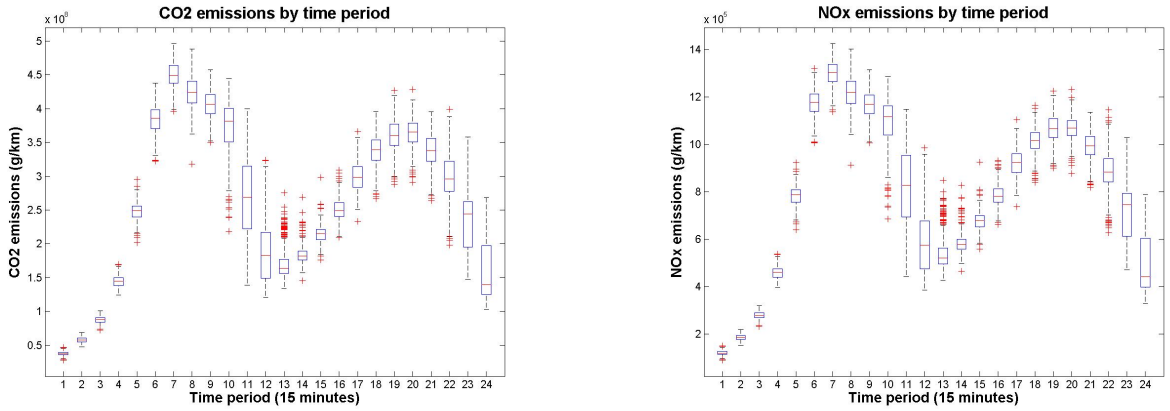
(b) Network spatial mean speed through time periods.

Fig. 2.16 – *Traffic variable evolutions through time periods: All simulations are represented at network level.*

Figure 2.16 ($a$) shows the distribution of traveled distance at network level through the time considering all traffic simulations; and in ($b$) shows the spatial mean speeds at network level. Analyzing both graphs, it is possible to identify some traffic states. The simulation starts with any cars in the network, so the vehicles move freely through the links from periods 1 to 6. This can be observed in ($a$) with the increase of traveled distances and in ($b$) with the spike of spatial mean speeds in the beginning of the simulations, which decreases along periods when vehicles rise on the network and the traffic demand increases.

Periods of congestion appear around the periods 8 and 21, when the spatial mean speed has a gradual decrease. In between, the network presents a free flow state, in a different way in comparison to first periods, because there are more passenger cars and interactions between them. It is interesting to observe that the simulations are highly variable for traveled distance and mainly for spatial mean speed which presents a heterogeneous congestion at about the 21$^{st}$ period. Even in the beginning of the simulation, the variable values are dispersed, and increase differently through time, which confirms that the dataset is diverse enough to perform sampling methods.

If both traffic variables are combined to calculate the evolution of pollutant emissions through time, they will present almost the same shape as traveled distance but with more marked peaks. Emissions are directly proportional to the traveled distances and the emission factors, depending on the spatial mean speed. The emission factors are inversely proportional to the spatial mean speed, which means that lower speeds produce the higher emission factors. It explains the marked peaks on pollutant emissions graphs. Figure 2.17 shows the dispersion of the pollutant emission values by time at network level considering all simulations performed.



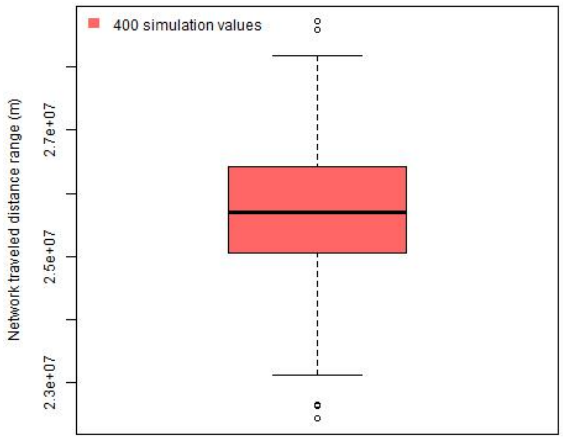

(a) Network $CO_2$ emissions through time periods.

(b) Network $NO_x$ emissions through time periods.

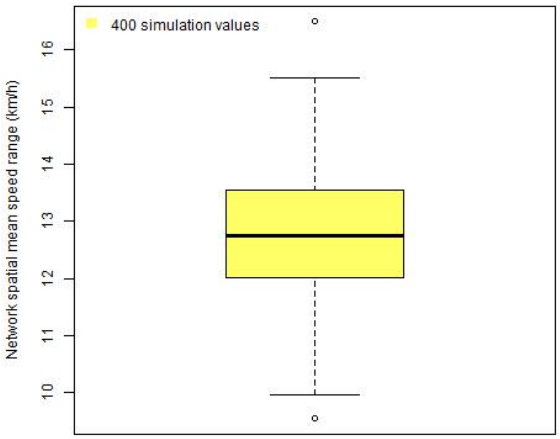
Fig. 2.17 – *Pollutant emissions through time periods: All simulations are represented.*

For both pollutants, the congested periods have higher peaks of emissions compared to free flow periods. Another characteristic is the dispersion of values. The periods which present more dispersed traffic variables present also more dispersed emissions values. This is explained by the heterogeneous behavior when the network starts to run out of congested state.

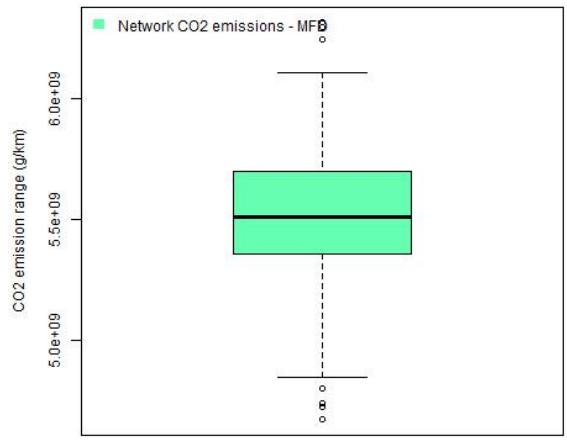
The existence of outliers was also analyzed from these simulations, which means that some simulations lie outside (i.e. is much smaller or larger than) most of the other values in a set of data. The aim is to observe how much the daily network values for each simulation are dispersed. Having scattered values is important to apply statistic methods without creating trends that may create samplings that do not represent the population. All variable values at the network level are represented in figure 2.18.



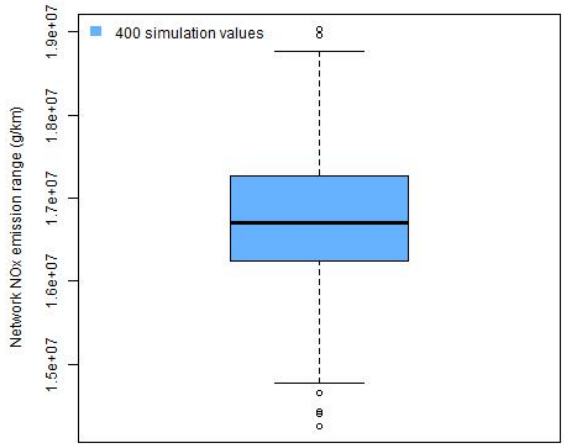

(a) Daily network traveled distance values.




(b) Daily network mean speed values.




(c) Daily network $CO_2$ emission values.




(d) Daily network $NO_x$ emission values.

Fig. 2.18 – *Distribution of daily network variable values from all simulations.*

Considering the values of all simulations, it is possible to observe the high variations of the network values of all variables. The traveled distance variable values are distributed between 23.000 km and 28.000 km of displacements inside the network. It presents 6 values outside of this range (i.e. outliers), the minimum value is at about 22.400km and the maximum reaches 28.600km of displacements.

The spatial mean speed also presents dispersed network values through simulations. The range of values go from 10 km/h to a little more than 15 km/h and has two outliers, for one simulation the network mean speed is around 9 km/h and for the other it is around 16 km/h. These low values at the network level are completely normal because: (i) the network is an urban area; (ii) the values are represented with space and time gathered and (iii) most periods are semi-congested or congested through the time in each simulation.

For the pollutant emissions, the values are highly variable as for traveled distance. The normal

distribution presents their outskirts around 185 and 235 kg of $CO_2$ emissions and 0,6 and 0,75 kg of daily $NO_x$ emissions. The $CO_2$ has 8 values outside of this range, the minimum value of emission is 182kg and the maximum will be around 293 kg. In the same way, the $NO_x$ emissions variable has 7 outliers with extreme values around 0,54 and 0,73 kg by journey.

In general, all the variables do not present a lot of extreme traffic behaviors (i.e. outliers), that means the simulations are consistent with few abnormal ones. Traveled distances and spatial mean speed outliers come from different simulations ID while emissions got all from both. Figure 2.19 shows the ID of simulations that represents outliers in the network distribution variable values.

| Outliers from travel production | Outliers from mean speed | Outliers from $CO_2$ emissions | Outliers from $NO_x$ emissions |
|---|---|---|---|
| Simulation ID | Simulation ID | Simulation ID | Simulation ID |
| 45 | 110 | 45 | 45 |
| 62 | 153 | 62 | 62 |
| 96 | | 96 | 96 |
| 219 | | 110 | 153 |
| 294 | | 153 | 219 |
| 374 | | 219 | 294 |
| | | 294 | 374 |
| | | 374 | |

Fig. 2.19 – *Outliers simulation ID by variable.*

It is interesting to observe, by analyzing the simulation ID that corresponds to outliers, the emission outliers are a mixture of outliers that come from traffic variables. It is explained by the strong correlation between traffic variables and emissions, mainly by the direct correlation between traveled distance and emissions. The last will be discussed in the next lines.

### Correlations between variables

The correlations between traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions were studied. The aim is to identify the existence of correlation and how strong they are, mainly between traffic variables and emissions. Each variable will be represented by the daily network values, meaning the variable values were aggregated in space and time for each simulation. The correlation between them at the network level is shown in figure 2.20.

The correlations were calculated with 95% of interval confidence in the significance test. The traveled distance is highly correlated with emission variables due to the fact that they are directly proportional and this can be seen over the time periods. It is interesting to observe also how the spatial mean speed is correlated with other variables through the time. In appendix A the correlation evolution is presented by time period. The spatial mean speeds present strong inverse correlation in periods 11, 12, 13, 23 and 24. These periods show an increase of the spatial mean speed and consequently a decrease of traveled distance and both pollutant emissions. Contradictorily, periods 7, 8 and 9 are the unique with a positive correlation between spatial mean speed and the other variables. These periods represent a strongly congested traffic state. The first periods do not have a correlation at all or just present weak ones. It can be explained by the fact that the network did not have vehicles loaded before the first period, consequently there are not enough passenger cars distributed over the network. To the other time periods, the spatial mean speed does not present a strong or significant correlation. To conclude, for all periods the traveled distance has a direct and strong correlation with both pollutant emissions and consequently a strong influence on them. The spatial mean speed has strong influence only over periods that present a free flow state, most of the time an inverse correlation.

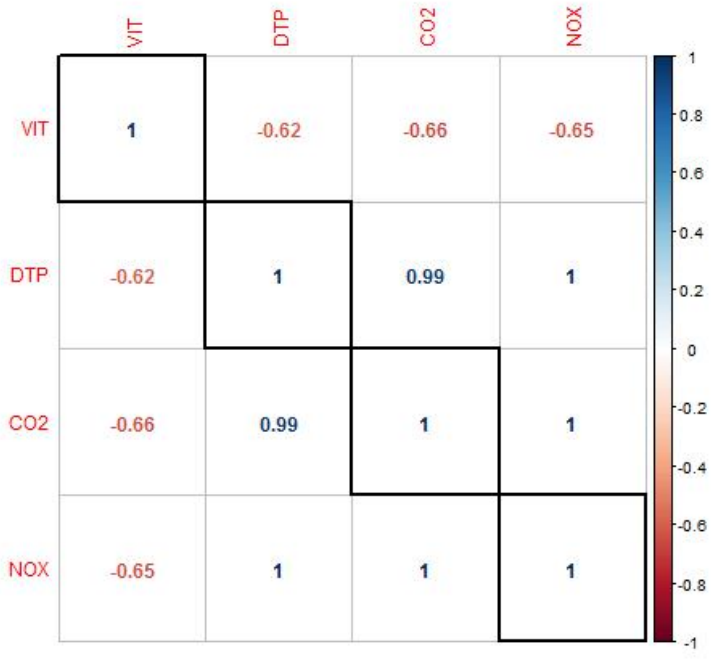
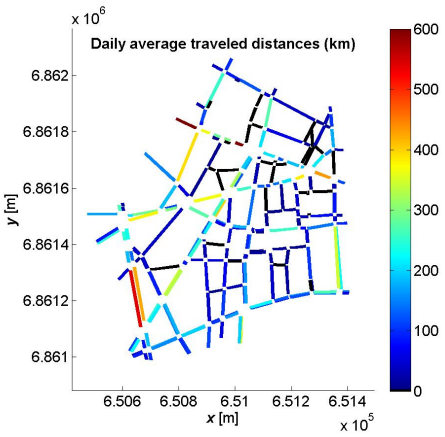
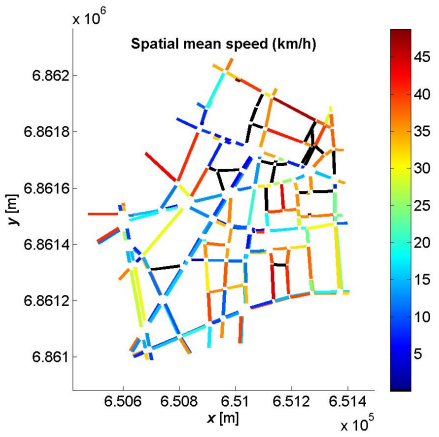



Fig. 2.20 – *Correlation between variables at network level.*

To understand how the variables spread through the network, each one was studied at link level. Figure 2.21 shows the mean values between all simulations (over a year of data) for a day's traffic.



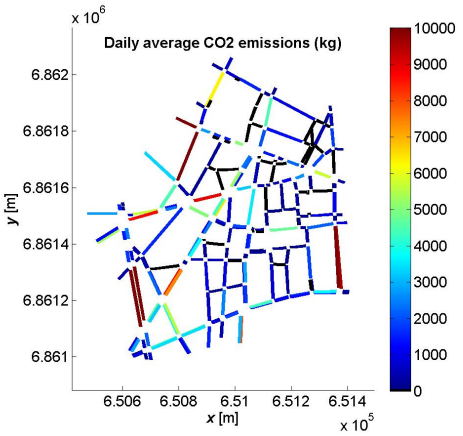

(a) Mean traveled distance values by link.



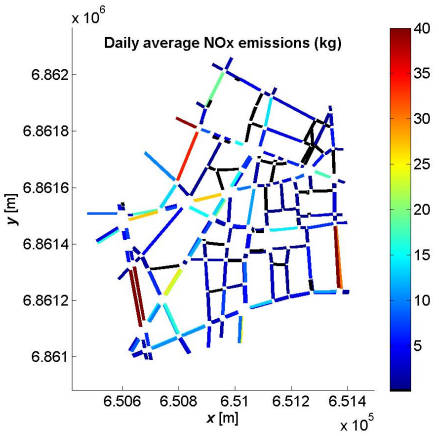

(b) Spatial mean speed values by link.

Considering 400 days of traffic data, it is possible to observe in (*b*) that the main axes have lower mean speeds, consequently a congested state. Analyzing the average traveled distance by link, some parts of the main axes have total distance traveled around 300 to 400 km a day. Sums up the displacements and the lower mean speeds, this same axis have higher emissions than the other ones, mainly the entry of cars by the Boulevard Saint-Germain as referenced as 13 in figure 2.2.

To have a clear view of each variable in each link, the 10% of the most traveled and most polluted links were ranked using the average value considering all simulations. This percentage rate represents a good one to compare with selected links by selection methods applied on this network. For spatial mean speed, the ranking was made considering links that have average speed below or equal to 10



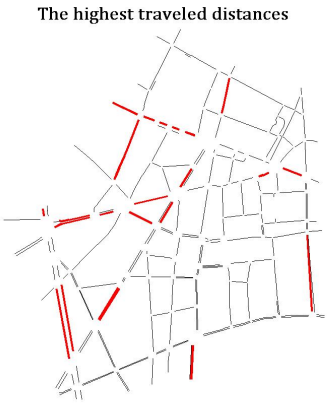

(c) Mean $CO_2$ emission values by link.



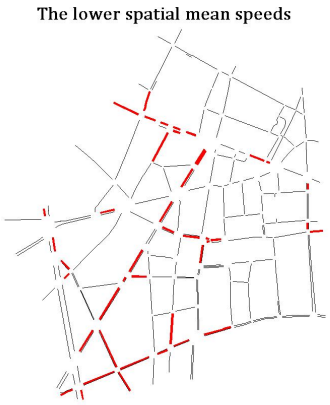

(d) Mean $NO_x$ emission values by link.

Fig. 2.21 – *Average values of variables by link considering more than a year of data.*

km/h and consequently higher emission factors. Figure 2.22 shows the 10% of links that had most displacements and emissions in average and also the links that have the spatial mean speed lower or equal to 10 km/h.



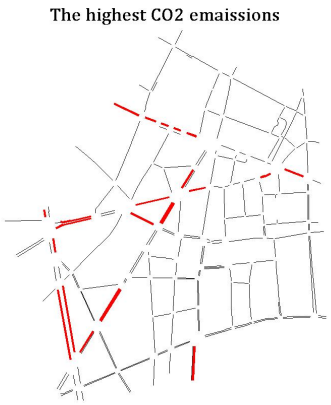

(a) 23 links most traveled in average.



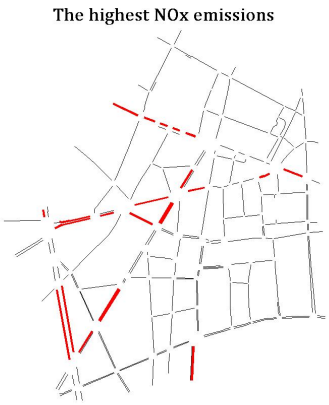

(b) Links with spatial mean speed below or equal to 10km/h.

Between these ranked links, it is possible to observe that the most traveled and polluting ones are almost the same, and this is because they are strongly correlated. For spatial mean speed, the most important is to identify the links that have average speed lower or equal to 10 km/h, because in this range of speed emission coefficients are higher. There are 44 links that present this characteristic. The ranked links of each variable were compared to identify the percentage of equal ones. Table 2.1 show the percentage of common links.

These ranked links represent, on average, 67% of the total traveled distances of the network per day simulation and the 23 most polluted links represent 62% of the total emissions. Most of these links are situated in major axes. Besides the network was split in two parts, the major and minor axes. In Paris 6$^{\text{th}}$ district study case, 30% of the links are situated in major axes, 3 times the size of the ranked links already studied. These links represent 58% of the total traveled distances per day in the network and 61% of the total emissions by $CO_2$ and $NO_x$. Furthermore, the fact that minor axes have a contribution around 60% and represent 70% of the network cannot be neglected. The 69 links




(c) 23 links most $CO_2$ polluter links on the network.




(d) 23 links most $NO_x$ polluter links on the network.

Fig. 2.22 – *10% of the most representative links for each variable in the network.*

| \ | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| DTP | \ | 34,8% | 82,6% | 87,0% |
| VIT | 34,8% | \ | 47,8% | 43,5% |
| $CO_2$ | 82,6% | 47,8% | \ | 95,7% |
| $NO_x$ | 95,7% | 43,5% | 95,7% | \ |

Tab. 2.1 – *Percentage of common links present in the variable rankings.*

that are situated in major axes for the 6$^{th}$district of Paris are shown in figure 2.23.

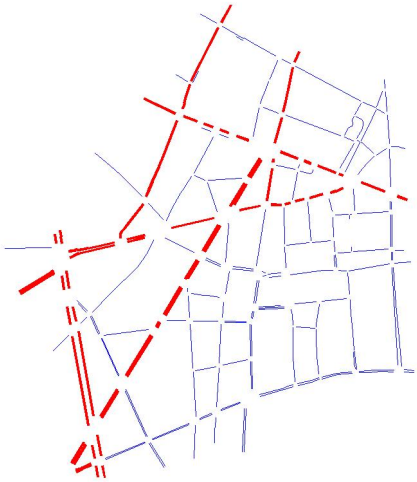
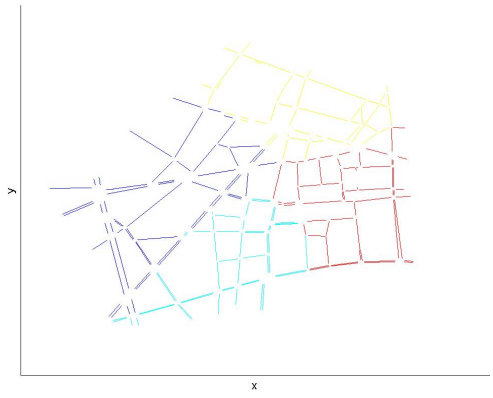


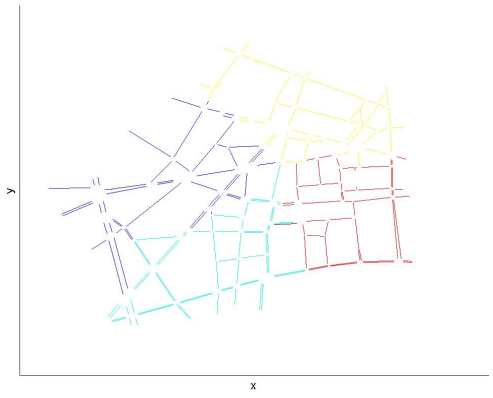
Fig. 2.23 – *Major axes highlighted in red and minor ones in blue.*

When a network is separated in groups of links that represent similar traffic characteristics it is called clustering. The spatial cluster identification can evaluate the geographic variation of active transportation and identify neighborhoods with unusually high/low levels of traffic.After having partitioned the network in two thanks to a subjective criterion (major-minor axes), the clustering method called "Snakes" (Ji and Geroliminis, 2012b) was used to partition spatially the network according to

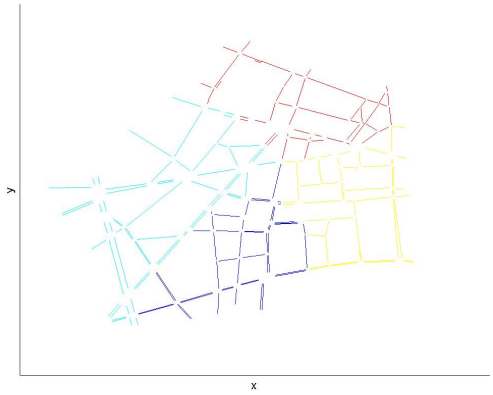
the variables studied. In this section, the method will be addressed only to show the partitioning of the network and identify groups of links and the contribution of each cluster in the total value of the variable analyzed. This method is explained in detail in Chapter 4. The network was partitioned in 4 clusters for each variable. Figure 2.24 shows the partitioning per variable.



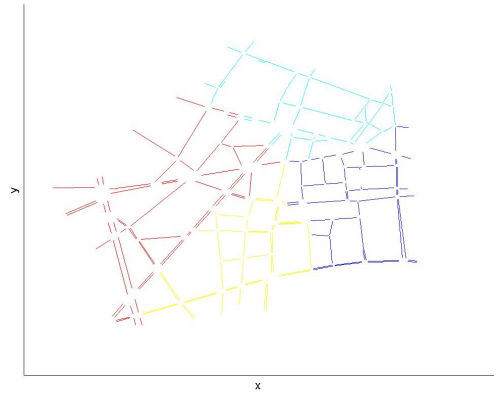

(a) Network partitioning take into account the travelled distances.




(b) Network partitioning take into account the spatial mean speed.




(c) Network partitioning take into account the $CO_2$ emissions.

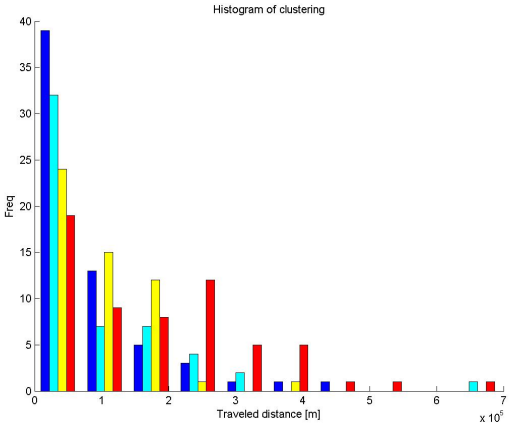



(d) Network partitioning take into account the $NO_x$ emissions.

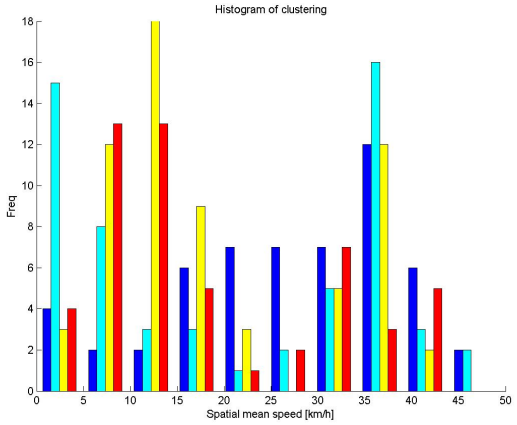
Fig. 2.24 – *Network partitioning by variable.*

As can be seen in figure 2.24, the partitioning between all variables are similar; each cluster has between 22% and 30% of the network links. Traveled distance and pollutant emissions partitioning conduct to 3 clusters representing 16 to 20% of the total and the last represents 42 to 47%. three clusters which the contribution value according to the total are between 16% and 20% each one and the last cluster that represents between 42% and 47%. The value distributions are represented in figure 2.25.

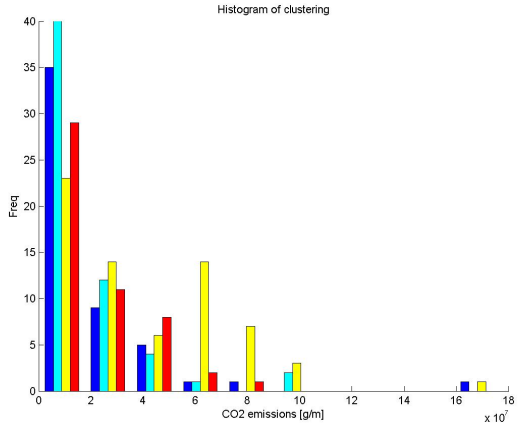
The most common values in each cluster are between 0 and 1 in the variable scale except for spatial mean speed. The mean speed shows a bimodal distribution and the linear variables a right-skewed one with a few outliers in some clusters. This partitioning will be used to apply statistical methods in the defined clusters, in order to estimate variables values by a region with more accuracy and less dispersion.



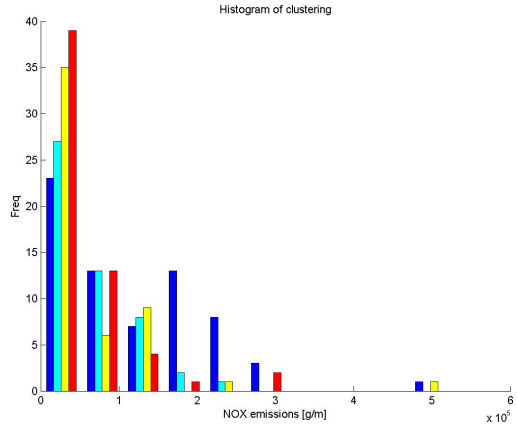

(a) Travelled distances.




(b) Spatial mean speed.




(c) $CO_2$ emissions.




(d) $NO_x$ emissions.

Fig. 2.25 – *Histogram of variables by clusters.*

## 2.4 Conclusion of Chapter 2

In this Chapter was presented the descriptive analysis of study case of the thesis using simulated data of more than a year of traffic information. The network and the simulation environment were described. The variable values, as traffic and pollutants emissions, were defined for two types of source information: the spatial and punctual sensors. A comparison between both was made to define the best variable description to use as the basis of the work relying on selection methods proposed in the next chapters. This study showed that traffic information from punctual sensors can provide biased values. The total traveled distance values are overestimated in 3% on average and this bias is accentuated in congested periods. Considering mean speed, the overestimation can reach more than 100% of error. In free-flow states this difference can be increased and reach 14 km/h de difference. Both traffic bias can induces an underestimation of the network emissions in average 14%. The traffic information from punctual sensors are usually used by city managers to apply new strategies or conducting studies on mobility and emission reductions. However, the precision and reliability of traffic information gave by the microscopic simulation is needed to estimate with accuracy the pollutant emissions and to study their variability according the spatial and temporal scales. So, the spatial data was retained to study and apply the sampling methods that are proposed through this thesis.

The network behavior was presented to understand the correlation between variables, major and minor axis identification, their contribution according to the total values and the most polluting zones. All that information will be used to support and analyze the results obtained with statistical methods or to improve them.

# 3

# Sampling link selection based on the LASSO method

To better assess the contribution of traffic to air pollution, detailed traffic data can be defined (i) experimentally or (ii) by simulations performed for a region under study. Regarding (i), one of the major difficulties in obtaining this traffic information is related to the high financial cost of implementing sensors. Regarding (ii), the complexity of the computer processing involved must be taken into account as it increases as a function of the scale of the network (i.e. the ratio between the number of links and the number of the observations). Sampling methods will be proposed to solve both problems. The aim of this work is focused on the representativeness of traffic data sampling to significantly reduce the volume of data to be processed while maintaining an accurate estimation of the overall results in terms of air pollution. Thus, it is necessary to define a representative sample by considering its evolution in time and space. For example, rather than using all the traffic information produced by a microscopic model, the goal is to define a subset of representative links and reference time periods from which it is possible to calculate emissions.

The applications of such techniques are numerous. In addition to significantly improving computing time, the development of appropriate sampling methods could also help identify key components of a network or travel types and thus improve assessments a posteriori (optimal positioning of measurement stations, definition of reference rounds for vehicles with embedded measurement instruments, etc.). The techniques covered by this thesis could also be useful for real-time assessments of air quality. Indeed, sensor networks for air pollution are generally sparse and do not specifically discriminate the contribution of road traffic.

We will use different traffic data sampling methods, taking into account the phenomenon of congestion and spatial and temporal information. The objective is to assess how relevant these methods are when calculating the solution for the total emissions, the mean network speed and travel production.

The amount of traffic data has vastly increased over time and nowadays thousands of descriptors can describe different aspects of traffic behavior that can be calculated by data processing and dedicated software. However, when modeling a particular property of traffic, it is reasonable to assume that only a small number of descriptors are actually correlated to the experimental response and are therefore relevant for building the mathematical model of interest. Consequently, a key step is the selection of the optimal subset of variables (*i.e.* traffic descriptors as loops or sensors) to develop the model. This is precisely the aim of the so-called variable selection method, which allows: (i) improving interpretability (simple models); (ii) avoiding the omission of significant effects, thus reducing noise; (iii) increasing the model's predictive ability; and finally (iv) accelerating modeling time.

Statistical methods for variable selection have evolved from the simplest to the most efficient tools

according to needs and scientific fields. Moreover, certain new methods have been proposed in the literature that can simultaneously combine regression and variable selection (Cassotti and Grisoni, 2011). The least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) is a recent statistical method that is mainly based on linear regression. In the last decade this method has attracted much attention because of its employability in many fields and its capacity to take into account many more variables than other statistical methods (Bien, 2016). Applying linear regression methods to large amounts of predictors can present some difficult aspects such as the goodness of the fitted model, which means the fitted model does not rely on the data observed; consequently over or under fitting can be expected. Also, interpreting the results obtained from linear models can be complex in this case. Based on this information, the methodology of LASSO overcomes this issue by selecting a small number of variables and use them to build a solid and trustworthy model (Bien, 2016).

In the transportation literature, the LASSO method is most used in traffic flow predictions. In (Li et al., 2015) the authors applied LASSO to quickly filter out most of the unrelated data that came from various sensors for Intelligent Transportation technology (ITS), maintaining both the temporal characteristics and spatial dependence of the original traffic flow time series. Most of the data used for real-short time traffic forecasting comprise non-linearity and suffer from potential data-induced collinearity. In (Kamarianakis et al., 2012) these difficulties were dealt with using the LASSO method. The problem of input as a feature of the spatial-temporal neighborhood in forecasting the value of space-time series at a given point in space and time using LASSO was studied by (Haworth and Cheng, 2011).

## 3.1 LASSO - Least absolute shrinkage and selection operator

LASSO was proposed by (Tibshirani, 1996) as a regression method that penalizes regression coefficients, by penalizing results in a situation where some of the coefficients are exactly equal to zero and thus reduce the number of predictors inside the model. The stronger the penalty, the more the coefficients can be shrunk towards zero.

Let us take a set of explanatory variables $(X)$ for $1 \leq n \leq p$, where $n$ is the set of observations of descriptors $p$, to explain a variable $Y$ linearly, but nothing ensures that all the variables involved are explanatory. So we have a set of potentially explanatory variables or candidates. Our goal is to identify the explanatory ones. Therefore, it is necessary to choose a model among $2^{\mathrm{p}}$ possibilities. Considering the latter, how can we choose the right model? It is not possible to study all the possibilities when $p$ is large; in addition, how is it possible to know which model is better than the other ones. In certain cases the LASSO method offers a solution to this problem. This is convenient when dealing with highly correlated predictors, where standard regression will usually have very large regression coefficients (Huang et al., 2006).

### 3.1.1 The idea underlying the method

We seek to explain a variable $Y$ linearly by $p$ variables that can be potentially explanatory of $X$. To this end, we simulate $n$ observations. The variable model $Y$ is described below:

$$Y = X\beta + \varepsilon \tag{3.1}$$

where, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is a vector of $n$ random variables with an average value equal to zero and a variance $\sigma^2$ which corresponds to the noise in the observations (*i.e.*which may contain all the explanatory variables not taken into account in the model); $Y \in \mathbb{R}$ is a vector that corresponds to $n$ observations of $Y$ (*i.e.* the response values that we want to predict). $X = (X_{.,1}, \cdots, X_{.,p}) = ((X_{1,.})^T, \cdot, (X_{n,.})^T)^T$ is a $n \times p$ matrix, where $n$ are the lines of the matrix and correspond to the observation values of the predictor $X_p$; $p$ is the columns of the matrix which correspond to the variables $X$. Mathematically speaking, $X_p$ is column p$^{\text{th}}$ that corresponds to the predictor p$^{\text{th}}$ of $X_p$. $X_n$ is the line n$^{\text{th}}$ that corresponds to the observation n$^{\text{th}}$. $\beta \in \mathbb{R}^p$ is the parameter that must be estimated and is indexed by $n$ to allow its coefficients and its size to vary as $n$ increases ($p$ may depend on $n$).

If $X_p$ variables are not all relevant, the goal is to eliminate the unnecessary variables and only them. The idea of LASSO is not to perform a classical linear regression but a regularized regression that makes some of $\beta$ coefficients equal to zero. This involves estimate $\lambda \in \overline{\mathbb{R}}_+$. Considering the latter, the $\beta$ coefficients are calculated as follows considering $\lambda \in \overline{\mathbb{R}}_+$.

$$\widehat{\beta}(\lambda) = \underbrace{argmin}_{\beta \in \mathbb{R}^p}(\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1) \tag{3.2}$$

where $\|x\|_2^2 = \sum_{i=1}^n x_i^2$ and $\|x\|_1 = \sum_{i=1}^p |x_i|$.

The parameter $\lambda \geq 0$ controls the power of regularization. If $\lambda = 0$, the LASSO method corresponds to a classical linear regression (if $p \geq n$). On the contrary, if $\lambda = \infty$, all $\widehat{\beta}(\infty)$ are equal to zero. Increasing $\lambda$ induces certain $\widehat{\beta}(\lambda)$ coefficients to decrease until exactly zero. The last model is equivalent to the following:

$$\widetilde{\beta}(t) = \underbrace{argmin}_{\beta, \|\beta\|_1 \leq t}(\|Y - X\beta\|_2^2) \tag{3.3}$$

considering for all $\lambda \in \overline{\mathbb{R}}_+$, $t \geq 0$ such that: $\widetilde{\beta}(t) = \widehat{\beta}(\lambda)$. Indeed, simply take $t = \|\widehat{\beta}(\lambda)\|_1$ then for all $\beta$ such that $\|\beta\|_1 \leq t$, $\lambda\|\beta\|_1 \leq \lambda\|\widehat{\beta}(\lambda)\|_1$ therefore, by defining $\widehat{\beta}(\lambda)$, $\|Y - X\beta\|_2^2 \geq \|Y - X\widehat{\beta}(\lambda)\|_2^2$.

This explanation allows us to understand intuitively why, in most cases, LASSO results in exactly zero for certain $\widehat{\beta}(\lambda)$ coefficients (Friedman et al., 2010). Figure 3.1 shows the case where LASSO gives exactly zero to the coordinates. The latter depend on the position of $A = \arg min_\beta(\|Y - X\beta\|_2^2$ in $\mathbb{R}$ and $t$. The smaller $t$, the more LASSO gives a zero value to the coordinates.

When the dimension expands into the area where $A$ gives exactly zero for at least one coordinate, this $A$ zone increases to become almost the entire $\mathbb{R}^p$ area.

The algorithm applied to optimize and solve LASSO uses a regularization path via Elastic Net (Zou and Hastie, 2005). (Ghosh, 2007) showed that this methodology using Elastic Net applied to real and simulated data is more efficient than the previous algorithm for solving LASSO developed by (Tibshirani, 1996), without prejudicing the final results due to a lack of representativeness. Also, (Zou and Hastie, 2005) highlighted that the elastic net may have a grouping effect which depends on the degree of correlation of the variables. This means that strong correlated variables can be in or out of the model together.

Elastic net uses the same type of data set as LASSO. Let us suppose that the data set has $n$ observations with $p$ predictors. Let $Y = (y_1, \cdots, y_n)^T$ be the response vector and $X = (x_1 \mid \cdots \mid x_p)$ be the model matrix, where $X_j = (x_{1j}, \cdots, x_{nj})^T$, $j = 1, \cdots, p$, are the predictors. A location and scale transformation are applied to the data, the response is considered centered and the predictors

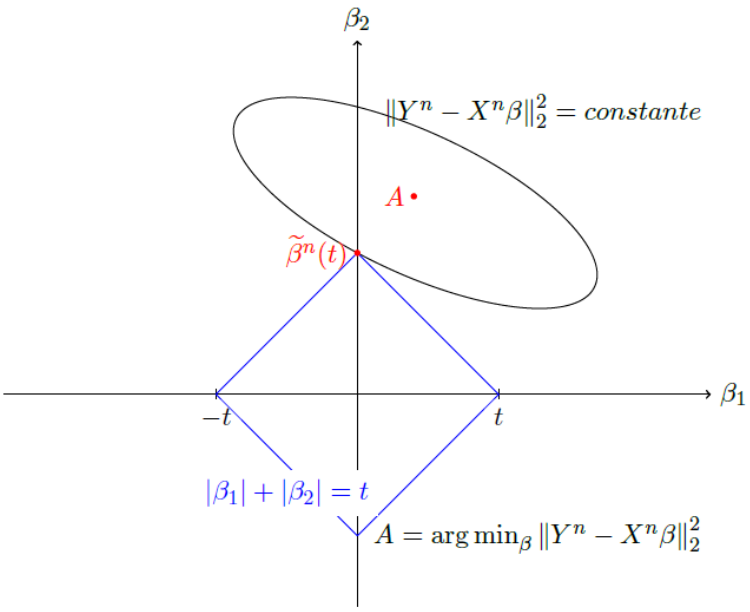



Fig. 3.1 – *Situation where A is inside the zone where $\beta_2$ is exactly zero (Friedman et al., 2010).*

are standardized, as shown in equations 3.4:

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0 \ and \ \sum_{i=1}^{n} {x_{ij}}^2 = 1 \quad for\ j = 1, 2, \cdots, p \tag{3.4}$$

Considering any $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, the elastic net criterion is defined by:

$$L(\lambda_1, \lambda_1, \beta) = \mid y - X\beta \mid^2 + \lambda_2 \mid \beta \mid^2 + \lambda_1 \mid \beta \mid_1 \tag{3.5}$$

where $\mid \beta \mid^2 = \sum_{j=1}^{p} \beta_j^2$ and $\mid \beta \mid_1 = \sum_{j=1}^{p} \mid \beta_j \mid$. The elastic net estimator $\widehat{\beta}$ is the minimizer of equation 3.5:

$$\widehat{\beta} = \underbrace{argmin}_{\beta} \mid y - X\beta \mid^2 \ \ subject\ to\ (1-\alpha) \mid \beta \mid_1 + \alpha \mid \beta \mid^2 \leq t\ for\ some\ t \tag{3.6}$$

Function $(1-\alpha) \mid \beta \mid_1 + \alpha \mid \beta \mid^2$ is called the elastic net penalty, which is a convex combination of LASSO and a ridge penalty. The elastic net is the same as LASSO when $\alpha = 1$. As $\alpha$ shrinks toward 0, the elastic net approaches ridge regression. In this work we consider $\alpha = 1$ to consider the LASSO approach in our dataset. More detailed explanations regarding the mathematical development are described by (Hastie et al., 2009a).

The next step is to determine the right $\lambda$ that can keep only true explanatory variables and eliminate others. The algorithm starts with a large value of $\lambda$ and runs the procedure until convergence, then $\lambda$ decreases using the previous solution as a warm start. One general approach to choosing the $\lambda$ value is to use a prediction error to guide the choice. Among the methods used is tenfold cross-validation.

Cross-validation is a method which starts by dividing the training data randomly into equal parts. Then, the learning method is applied to fit a model for the range of values of $\lambda$ in 90% of the data corresponding to one part. Finally, the prediction errors are calculated for the $\lambda$ range values in 10% of the remaining data that correspond to one part. This process is completed when the model is fitted for all the parts for each $\lambda$ value and the prediction error has been calculated. Then, all the

prediction error estimates are averaged (Friedman et al., 2010). Using the results of this process, it is possible to obtain an estimated prediction error curve as a function of the $\lambda$ parameter. It is always necessary to divide the data into training and validation sets. Cross-validation is always applied to the training set, since selecting the $\lambda$ is part of the training process while $\lambda$ determines the shrinkage process. The purpose of the validation set is to assess the performance of the selected model based on the estimation error.

A large number of algorithms for solving the convergence and optimization problem were tested, such as Convex Optimization in R (Koenker and Mizera, 2014), the Shooting Algorithm (Pendse, 2011), the Least-Angle Regression algorithm (namely LARS) (Efron et al., 2004) using the package developed by (Hastie and Efron, 2009), the Consistent algorithm to solve LASSO (De Vito and Veronica Umanità, 2011) and finally pathwise coordinate-wise descent optimization in Elastic Net. Apart from the last method, all the other algorithms gave a convex solution for our datasets. The results in this work came from the Pathwise Coordination Descent Optimization algorithm (Friedman et al., 2010).

The optimization method chosen computes an entire solution path in any $\lambda$ value, allowing the user to select a particular solution (*i.e.* model selection) from the ensemble. The algorithm that contains the optimization method is called *glmnet* and was developed by (Hastie et al., 2009b) and (Jiang, 2009). It is possible to evaluate the prediction performance at each value of $\lambda$, and also the corresponding model size (*the number of selected links*) from the mean-square prediction error in each $\lambda$ value. All this information can be evaluated easily from the prediction error graph. On the same plot, there are two highlighted models, one line that corresponds to the minimum prediction error among all the possible $\lambda$ values and to a model with $p$ number of predictors. The second line corresponds to the "one-standard error rule" defined by (Hastie et al., 2009b). This method consists in determining the biggest lambda value with a prediction error within one standard-error of the lambda model with the lowest number of errors which corresponds to the first line. These two highlighted models represent the best performance among the other possibilities, but the biggest difference between them is the number of selected predictors that each contains. The "one standard error rule" gives a model with fewer predictors than the model defined by the minimum error. The comparison of the estimation errors of both models show they are similar. For our purpose, the model using the "one standard error" rule is the best choice because it has a minimal number of selected regressors. The comparison between both models, performed to validate the choice made, is explained in detail in the next sections.

## 3.2 Datasets

Three types of datasets were built to characterize the dynamic behavior of the network. The goal is to answer several questions and identify the most relevant links on the network. The regressors are the variable values associated with the links. The sampling methods are applied to the links (spatial dependence) and the temporal aspect (dynamic characteristics) will be considered differently through the datasets. The dataset structures are explained below.

### 3.2.1 Static/static dataset

The first dataset, called static, considers only the daily traffic values for each link in the network, *i.e.* daily traveled distance, mean speed and emissions.

Each observation value from each link of the network was provided by a simulation. In the model, the regressors are the links and their observations are the total traveled distance, the mean speed and

the total emission values ($CO_2$ and $NO_x$) for each simulation (which represent the 6 most relevant hours of the day). Also, the proper calculation of the daily mean speed for a link is based on the mean speed in each time period and the total traveled times, see equation 2.3 in chapter 2.

Mathematically speaking, the variables such as traveled distance, mean speed and pollutant emissions were represented by the $X_{(n\times p)}$ matrix. The $p^{th}$ column corresponds to the predictor $p^{th}$ of $X_{(n\times p)}$. For all the variables in the static dataset, $p^{th}$ corresponds to the links of the network and each column $p^{th}$ is represented by a singular link. The network has 230 links, consequently the column length is equal to the number of links in the network. The $n^{th}$ line corresponds to the observation value. The observations can be represented by the total traveled distance, or the spatial mean speed or the total emissions for each pollutant. Each $n^{th}$ line corresponds to the daily value of the $p^{th}$ links. Each $n^{th}$ line is associated with a $n^{th}$ simulation. As explained before, 400 simulations were launched to produce data representing more than a year of traffic in the network. Thus, the length of the line corresponds to the number of observations.

The selection method was applied in these matrices based on vector $Y$ with $n^{th}$ lines. Vector $Y$ represents the daily network values (all links gathered), which means that each $n^{th}$ line can be represented according to the variable: the total traveled distance in the network; spatial mean speed; total emission values of the network for each pollutant.

The aim of this dataset is to estimate at network level the amount of emissions per day using only the daily traffic and emissions from a few links of the network. This method is proposed as a simplification of the process to evaluate the emissions in the network using only daily network traffic and emission values from the set of links selected.

### 3.2.2 Dynamic/dynamic dataset

The second dataset that will be studied is called dynamic and considers that each link comprises traffic data described every 15 minutes within the 6 hours of simulation. The data structure was built to obtain links and their time periods as regressors at the same time. The selection methods will be applied to the link and the time period. The aim is to identify which time periods are the most relevant for each link. This dataset allows estimating the network daily values using the traffic and emission data from links in a temporal range of 15 minutes. Thus the objective is to estimate daily values at network scale from the measure of 15 minutes of traffic or emissions.

This dataset is structured to allow the selection of links and their most relevant periods. $X_p$ is represented by links and periods of time. Each $p^{th}$ corresponds to a singular link and singular time period. The $X_n$ are the observations that correspond to the variable values of $X_{(n\times p)}$ from the $n^{th}$ simulation. The matrices have 5520 possibilities of predictors, *i.e.* $p$ (one predictor represents a link and period of time) with 400 observations each, *i.e.*. $n$. The selection method was applied to these sets based on the $Y$ vector that represents the daily network values, meaning the same vector as in the static datasets. Both datasets provide the network values as results, meaning spatial and temporal values considering the entire network and daily traffic. The goal is to further study the spatial-temporal correlations. In this case not only a link is selected, but also the relevant time period. In addition, the selected time period can provide an idea of traffic states, the level of congestion and observations of which characteristics are the most important for estimating network values.

### 3.2.3 Static/dynamic dataset

Static/dynamic datasets were built to consider that each link comprises traffic data described every 15 minutes within the 6 hours of simulations. The regressors are links and each observation corresponds to the traffic data for a period of time. A simulation is divided into 24 periods of 15 minutes and provides traffic data for each link. The observations for each link are composed for a time period with over 400 simulations. The particularity of this dataset is that the $Y$ values are now the variables for the global network but in a single time period (15 minutes).

As explained in the previous dataset, the $X_p$ also represents the links in this dataset. Each $p^{\text{th}}$ is represented by a link. Each $X_n$ corresponds to variable values from a time period. Each simulation provides 24 periods of time with traffic data (traveled distance or spatial mean speed) or pollutant emissions for each link of the network. Simulations were launched 400 times considering traffic inside the network with different levels of demand in each simulation. Each link of the matrices has 9600 observations. The selection method was applied to these matrices to select the links based on the $Y$ vectors with the same number of $n^{\text{th}}$ lines. Each line represents the network variable values of a period of time. The difference with the previous datasets is the temporal scale. Instead of estimating the network values for a day (i.e. the most relevant 6 hours) there will be an estimation for 15 minutes. Using the selected links its structure allows estimating, for example, the total network emissions for each 15-minute period and can be used to forecast emissions in the network for real-time tracking. Using the data grouped by blocks of 15 minutes and considered as variables allows integrating congestion in a model and thus performing better emission estimations. This issue will be discussed in the next sections.

## 3.3 LASSO method applied to the three different datasets

LASSO was applied to each variable in all the datasets and the results are described below. The repeatability of the method was studied separately in appendix B.1.

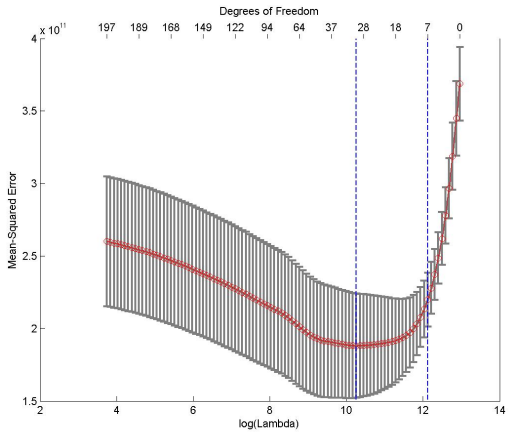
### 3.3.1 LASSO applied to the static/static dataset

#### Models proposed by LASSO

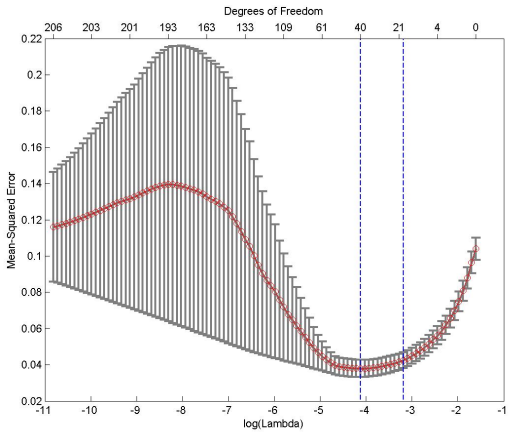
The explanations for the dataset are described in 3.2.1. This dataset has 4 different variables: total traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions. These variables are structured as $n \times p$ matrix, where $p$ are the links represented in the network while $n$ are the observations values for each link. Each link has 400 observations and they were split up randomly into two parts: the first represents 2/3 of the matrix and was used as a training set where LASSO was applied; the second part, 1/3 of the original matrix, represents the validation set for which the LASSO result selection will be validated and the associated errors will be quantified.

As explained in the section 3.1.1, LASSO highlights the two optimal models: one that considers the minimum error on the MSE curve (mean square error curve) and the second that refers to the model with the largest $\lambda$ values within one standard error from the minimum. The comparisons were made for each variable. The mean-square prediction error curves for traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions are shown in 3.2.

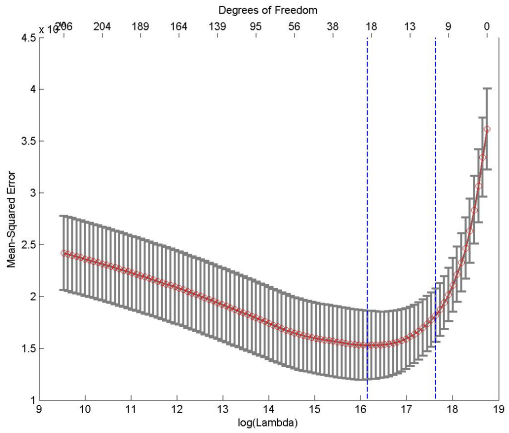
As can be seen in figure 3.2, each curve is plotted as a function of the corresponding ($\lambda$) parameter. The horizontal axis describes the model's complexity which increases as we move from right to left. The



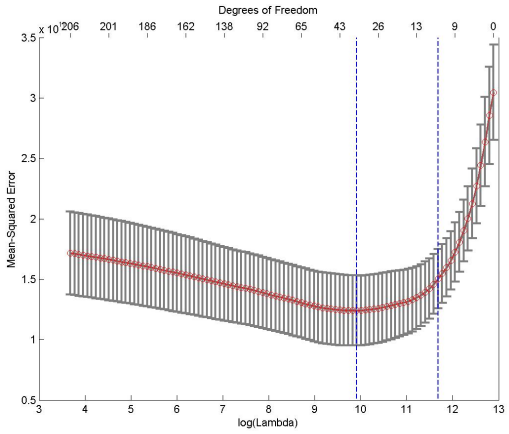

(a) Traveled distance mean-square prediction error curve.




(b) Spatial mean speed mean-square prediction error curve




(c) $CO_2$ emissions mean-square prediction error curve




(d) $NO_x$ emissions mean-square prediction error curve

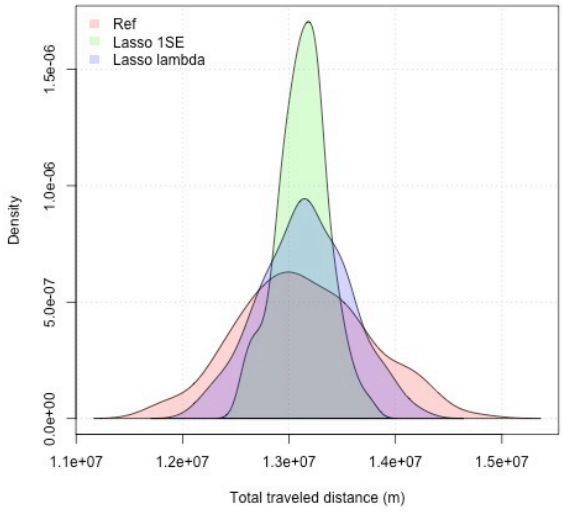
Fig. 3.2 – *Estimated prediction error curves for the variables in static/static datasets.*

estimates of the prediction errors and their standard errors were obtained by tenfold cross-validation, *i.e.* the training sets were divided into ten equal parts to apply the cross-validation method. The optimal models are indicated by the vertical lines on the plot. The top of each plot is annotated with the size of the models.

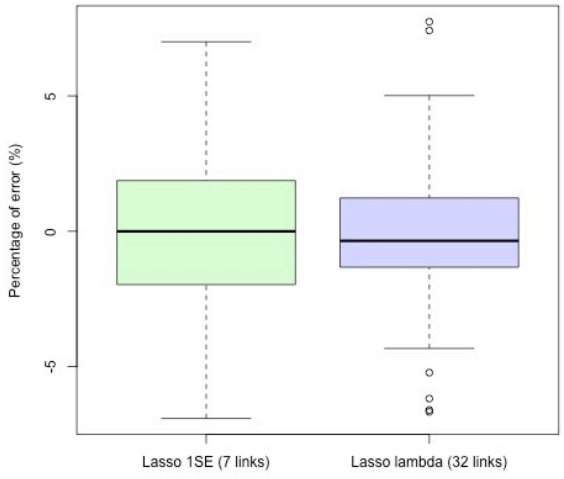
For our purpose, a model with a minimum number of internal predictors (links) is more interesting. To confirm the possibility of the choice, the predicted values of each model proposed by LASSO will be compared with the reference values and their relative errors will be calculated. Figure 3.3 shows the resulting values from both models for each variable and they are compared with the reference values (the original values that must be predicted).

The distribution of the lambda model is that most similar to the reference values, compared to the one standard-error lambda model. Also the error distribution of the lambda model is less spread out than that of the 1SE model. These facts can be explained by the number of selected links inside each model. The lambda models have 70% more predictors than the 1SE model considering all the variables. For example, for the total traveled distance LASSO selected 7 links in the 1SE model versus 32 in the lambda model. The same occurs for all the variables and can be observed for each one in 3.3 $(b)$, $(d)$, $(f)$ and $(h)$.

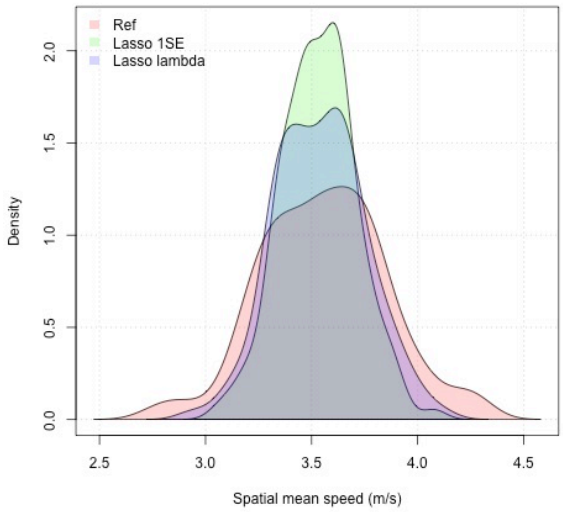
The 1SE lambda values do not have larger errors than the lambda model: in both cases 50% of



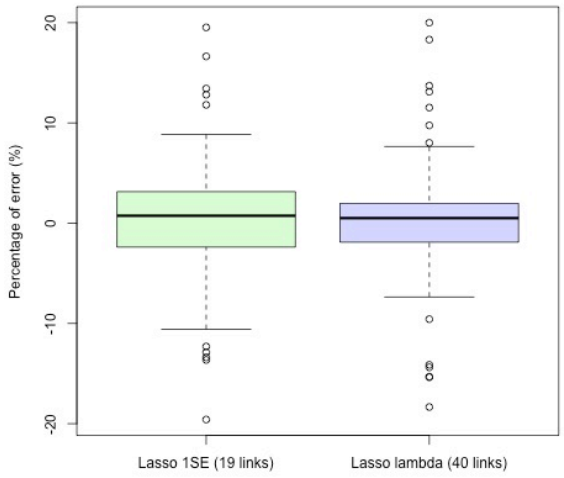

(a) Total traveled distance (DTP) density values.




(b) Total traveled distance error distribution.



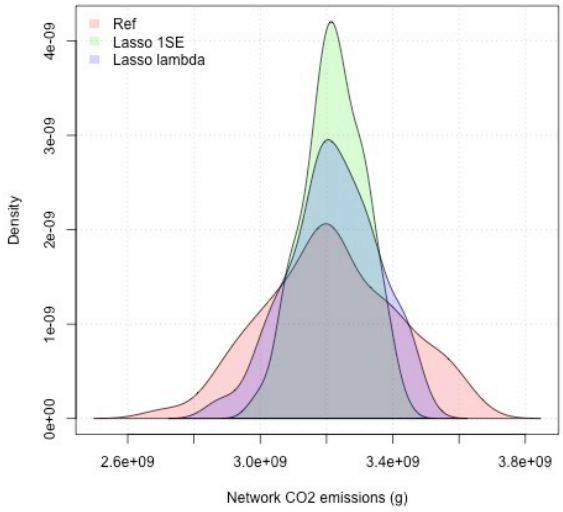

(c) Spatial mean speed density.



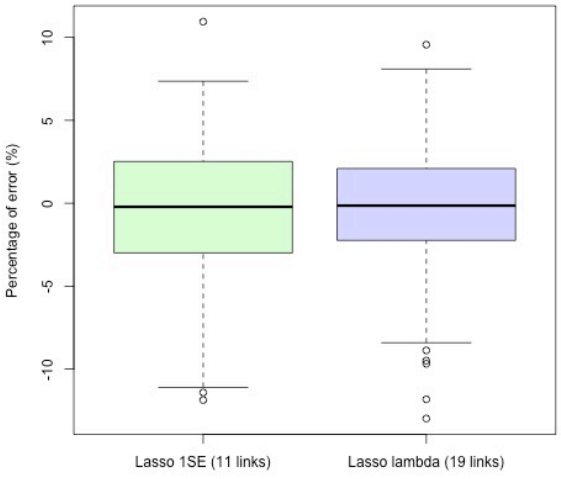

(d) Spatial mean speed error distribution.

the data with less than ±5% error for all the variables. Taking into account that both models have satisfying error distributions, the 1SE model appears more interesting for the study as it has far fewer predictors than the lambda model. The 1SE model needs from 3% to 5% of the network links to estimate all the variables considered with a total error distribution lower than ±10%.

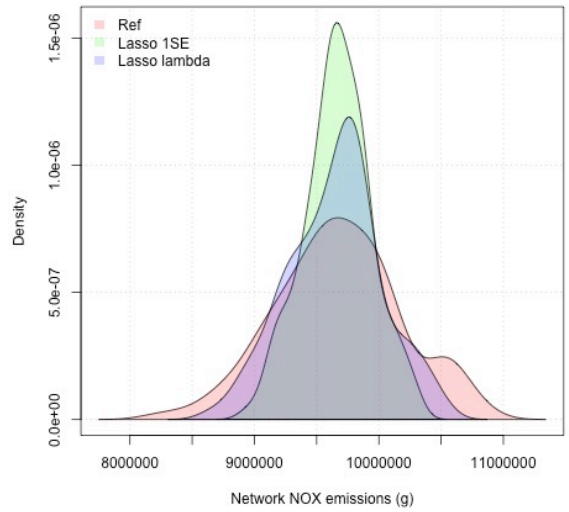
The dataset structure was built to simplify the process of estimating daily network values from the daily values of the links. At present, the traffic data of all the links of a network can only be obtained from microscopic simulation. Operationally, to facilitate the daily evaluation of the network values, the most important thing is to use the lowest possible number of links. A further study with 1SE models will be conducted in the next section, as it is smaller and has a reasonable error distribution in comparison to the lambda model.



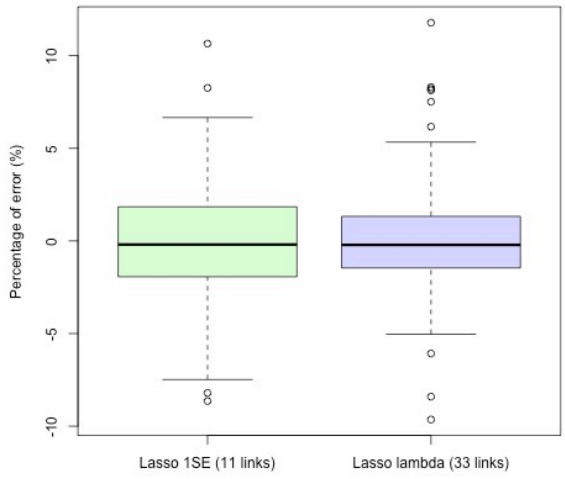

(e) Network CO2 emission density.




(f) Network CO2 emission error distribution.




(g) Network NOX emission density.




(h) Network NOX emission error distribution.

Fig. 3.3 – *Comparison between the predicted values of each model proposed by LASSO and the respective reference values.*

### The 1SE model

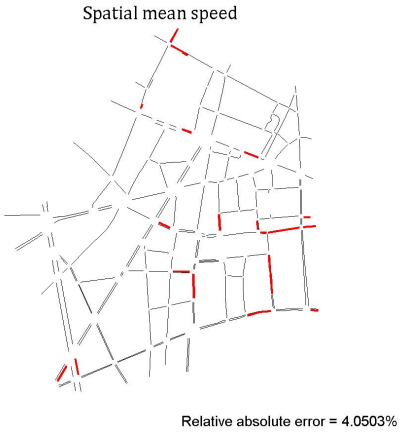
The model proposed by $\lambda$ with one standard error from the minimum square error (1SE lambda model) is the only model considered in this section due to its small size and because it is considered as one of the optimal models proposed by the LASSO method. The main aim is to analyze the selected links, the associated errors and the representativeness of the model built.

LASSO selected over 230 links present in the network, and for each variable studied it selected 7 links for traveled distance, explaining 43% of the data with a confidence interval of 95%; 19 links for spatial mean speed with 64% of the data explained by the model; 11 links for both pollutant emissions with a model that explains 54% of the data for $CO_2$ and 55% for $NO_x$ emissions. The relative errors

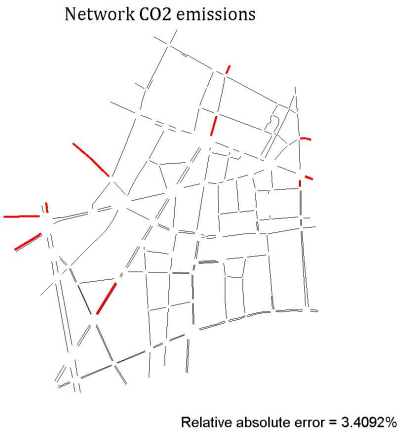
were calculated by comparing the results (values predicted by the model established using LASSO) with the reference values ($Y$) and are presented in figure 3.3. The Links selected for each variable are displayed in figure 3.4. Figure 3.5 shows the distribution of errors for each variable in the static dataset.



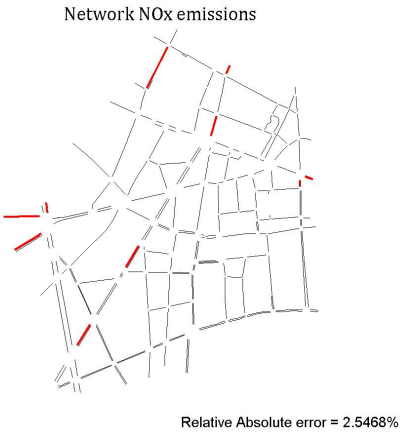

(a) The 7 links selected by the model for Total traveled distance.




(b) The 19 links selected by the model for spatial mean speed.




(c) The 11 links selected by the model for network $CO_2$ emissions.




(d) The 11 links selected by the model for network $NO_x$ emissions.

Fig. 3.4 – *Selected links by the $\lambda$ calculated with one standard error rule (1SE) and the mean relative absolute error of each variable.*

The links selected for each variable are completely different. The traveled distance selection has 4 links similar to the $CO_2$ selection and 3 similar to the $NO_x$ emission selection. The $CO_2$ and $NO_x$ link selections have 8 common links for a total of 11. Table 3.1 compares the common links for the four variable selections. It can be seen that the traveled distance and spatial mean speed variables have no common links. This can be explained by their opposing behaviors: traveled distance is a linear variable over the links whereas spatial mean speed is not. Considering the emissions, both present 72.7% of common links selected by their models, showing their strong correlation.

We compared the selected links and the most contributive link for each variable identified in chapter 2. This shows that 2/7 selected links are part of the most traveled links on the network considering the traveled distance variable; only 1/11 selected links on both pollutants emissions is part of the link with the highest emission level, and only 4 links out of 19 selected links have spatial mean speeds

| VARIABLES → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| MODELS ↓ | Validation set | Validation set | Validation set | Validation set |
| DTP | 100% | 0% | 57,1% | 42,8% |
| VIT | 0% | 100% | 0% | 0% |
| $CO_2$ | 36,4% | 0% | 100% | 72,7% |
| $NO_x$ | 27,3% | 0% | 72,7% | 100% |

Tab. 3.1 – *Ratio of common selected links between variables.*

lower than 10 km/h. Thus, it is possible to conclude that the selection was not based on their highest values.

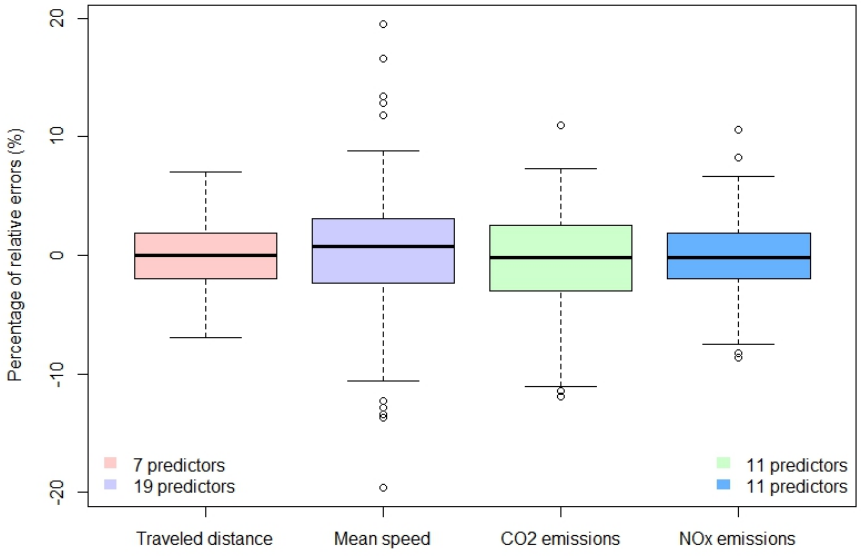



Fig. 3.5 – *Percentage of error distribution between the predicted variables values by the model built with selected links and the reference values of variables (Y).*

The relative errors were calculated by comparing the results (values predicted by the model established) with the reference values ($Y$). Figure 3.5 shows the distribution of associated errors in each variable for the static dataset.

All the variables have small average errors when considering models that have fewer than 9% of the links of the network selected. More than 50% of the data have errors lower than $\pm 5\%$ and, when considering all the data, the errors reach a little more than $\pm 10\%$ at most for all the variables. The 1SE lambda LASSO model can perform an optimal selection of network links and provide an estimation of daily traffic and emission values with reasonable errors using less than 1/10 of the data based on these results.

**Model size and error distribution**

The LASSO methodology allows building optimized models for each $\lambda$ value from the cross-validation method. Each $\lambda$ corresponds to a model size. For each variable, it is necessary to study which $\lambda$ model is the most optimized by considering their error dispersion as a criterion. The aim is to study how the error evolves through models with fewer selected links than the 1SE lambda (< 1SE model). We also study whether the number of selected links increases if the quality of the estimation also increases (> 1SE model). All the plots of traffic variables can be found in the appendix B.2 and the conclusions are given in the following paragraphs.

For the traveled distance, the analysis will start by investigating the error evolution compared to the lambda and 1SE lambda LASSO models. Figure B.5 in the appendix B.2 shows the possible size of the model with fewer predictors than the optimal lambda models.

The model with one star represents the 1SE lambda while the one with 2 stars is the lambda. If a error range of ±5% is fixed as the criterion for choosing the model, only the model with 32 selected links will be chosen. It should be borne in mind that a model with 1 predictor is insufficient to control and estimate variables from a transportation network, even if this model presents an error distribution of about ±10%. To better assess the emissions, it is necessary to know the strategic links distributed on the network. A single predictor can give only local information and cannot represent the evolution of the network variable. As the model with 32 selected links has an error distribution lower than ±5%, the models with fewer predictors than the lambda model are investigated in figure B.6.

In this case, models composed of from 27 links up to the lambda model are inside the range established, meaning that a model with 25 predictors can produce the same error distribution as the lambda model with 32 selected links. Models with more than 32 selected predictors are also studied. Figure B.7 shows the error distribution of these models. The model with 33 predictors presents the lowest dispersion of all the models studied. When the number of predictors starts increasing at this point, then the error distribution also increases. What is most striking is that having more predictors does not mean that the assessment will be of better quality. This will be more obvious when studying the other variables.

For the spatial mean speed, the lambda model (**) selected 40 links of the network while the 1SE lambda (*) model only selected 19. Figure B.8 shows the error distributions of all the modeling options with fewer predictors than the lambda model.

As shown in the figure, even the models selected by LASSO do not present an error distribution of ±5%. This condition will be investigated in models that have more predictors than the optimized models. Considering the error condition, none of these models are capable of estimating spatial mean speed with the ±5% error range. The least dispersed model is that with 31 selected links with an error range lower than ±8%. From this point, models with between 31 and 58 predictors have a similar error distribution, after which the error starts to increase. For both traffic variables, the higher the number of links inside does not lead to an increase in the quality of the estimation. Using a linear method, the spatial mean speed continues to be difficult to estimate with reasonable errors, compared with the traveled distance.

The same analysis will be performed for emissions. Figure 3.6 shows the different models until the number of predictors reaches that of the lambda model for $CO_2$ emissions.

The lambda model selected 19 links while the 1SE lambda selected only 11. Neither of these models has an error distribution of around ±5%. The models that had more predictors than the lambda model were investigated for spatial mean speed. Figure 3.7 shows models that can include up

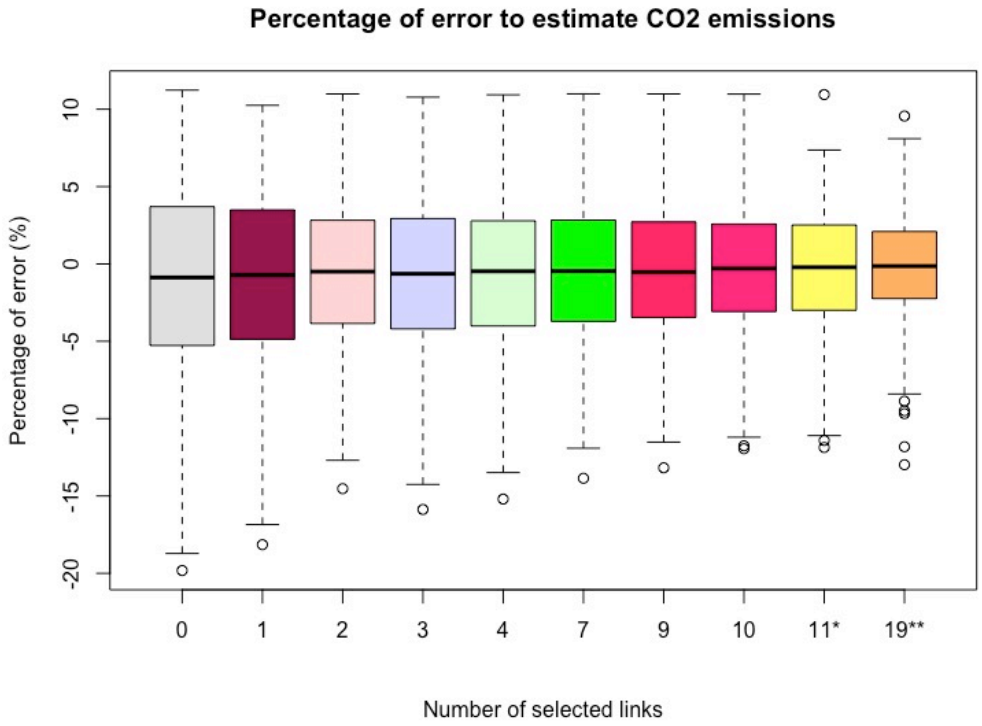



Fig. 3.6 – *The percentage error of models with fewer predictors than those of optimized lambdas.*

to 95 predictors.

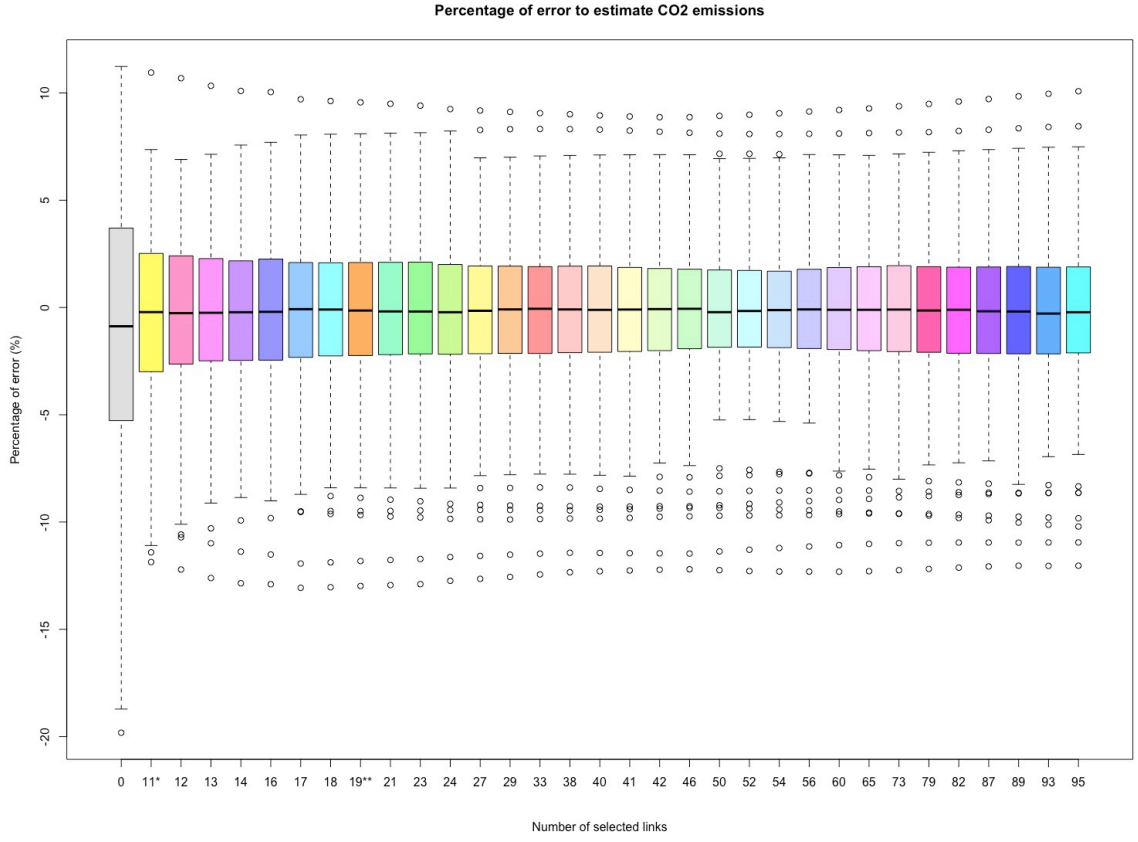



Fig. 3.7 – *The percentage errors of models with more predictors than optimized lambdas.*

It is noteworthy that no models have an error range lower than ±5%. The model with 50 selected

links, which represents 22% of the network, is the model with the least dispersion. The models with 52 and 54 selected links have similar error distributions.

The last variable studied is the pollutant $NO_x$. Figure 3.8 shows the models with the fewest predictors and figure 3.9 the models with more predictors than the lambda model.

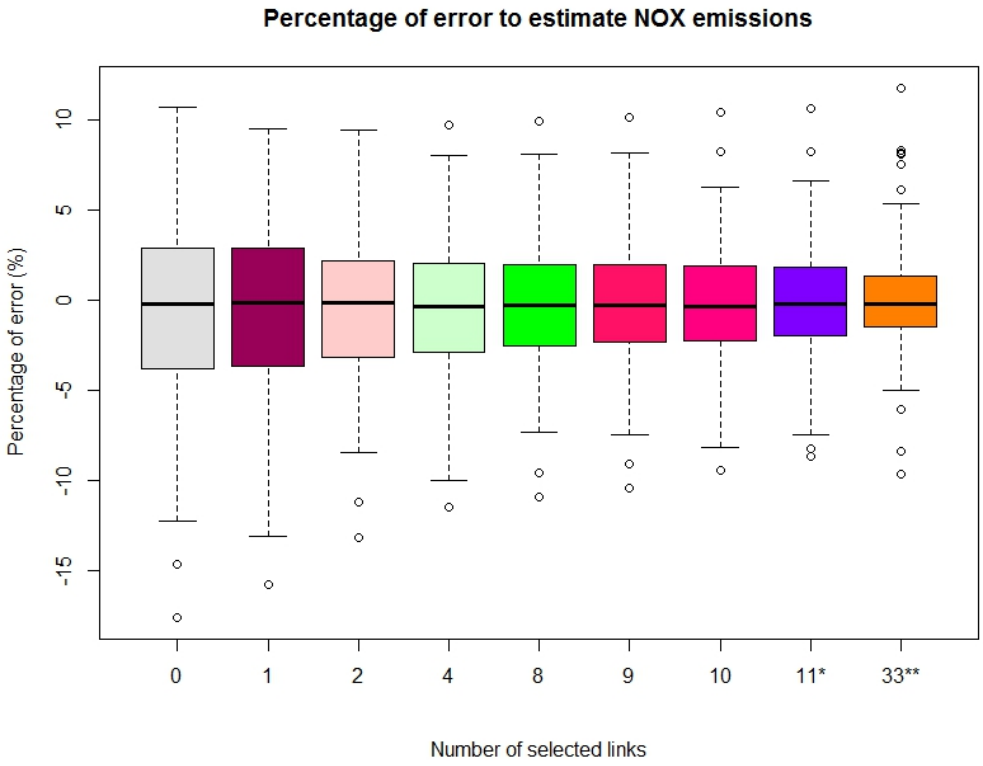



Fig. 3.8 – *The percentage errors of models with fewer predictors than optimized lambdas.*

The same phenomenon studied for $CO_2$ can be observed for the $NO_x$: the lambda model with 33 selected links has a percentage error of slightly over ±5%. Thus models with more than 33 predictors were also studied. The error distribution of models with between 36 and 44 selected links is in the range of ±5%. Between them, the models with 39 links have the smallest dispersion of all the models.

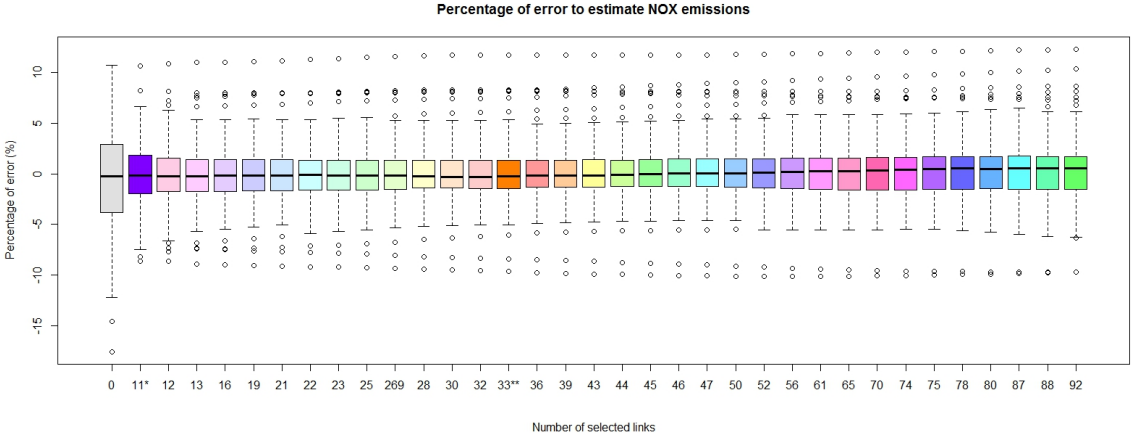



Fig. 3.9 – *The percentage errors of models with more predictors than optimized lambdas.*

Considering all the variables and the error range condition, the models used to estimate traffic data require about 14% of links from this network in order to ensure that the estimation is less dispersed.

For the emissions, 22% of the network is necessary to estimate $CO_2$ emissions and 17% for $NO_x$ emissions. These percentages are probably not realistic for operational applications. If we consider the criterion as the number of selected links, the 1SE lambda model is the best choice.

### The robustness of the models

A cross-analysis was performed to determine whether one of the 4 models built by 1 SE lambda for each variable could be used to determine other variable values in view to obtaining a set of selected links that can be used to quantify the network values for all the variables. Table 3.2 shows the average absolute errors of the model built based on the variable represented in the lines and applied to the variable values featuring in the columns.

| Daily average percentage of error (absolute error) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 7 | 3,04% | **2,70%** | **2,47%** | 6,29% | 5,87% | 2,39% | 2,73% | 2,40% | 1,92% |
| VIT | 19 | 8,26% | 2,08% | 1,97% | **4,12%** | **4,05%** | 2,36% | 2,69% | 2,32% | 2,22% |
| $CO_2$ | 11 | 4,78% | 1,95% | 1,63% | 6,21% | 5,85% | **3,01%** | **3,41%** | 2,31% | 1,93% |
| $NO_x$ | 11 | 4,78% | 1,93% | 1,65% | 5,38% | 5,26% | 2,23% | 2,82% | **2,87%** | **2,55%** |
| Number of observations | | | *267* | *133* | *267* | *133* | *267* | *133* | *267* | *133* |

Tab. 3.2 – *The average absolute error of the model established for one variable and applied to another.*

The table shows the average absolute error in the training and validation sets of the selected links of one variable applied to another and the average error for LASSO applied to the variables (red values). The same training set of each variable was used and linear regression was performed on the links selected by the LASSO model. The objective was to find the coefficient ($\beta$) values of each selected link adapted to the variable under study. In general, for all cases, the average absolute error values remain in the same range as the original LASSO method.

In addition, the errors between the training set and the validation set were compared. It can be seen that in some cases the average error is smaller in a validation set than in a training set. Mathematically speaking, the errors in the training set should be smaller because the models were constructed based on the training set data. To better evaluate the performance of the prediction, the mean square error will be considered. The reason for using a squared difference to measure the "loss" between the training and validation sets is consistency with the regression objective function. The variance and the square bias components can be associated with an estimator precision (small variance) and its accuracy (small bias), by averaging the squared difference over the distribution. Table 3.3 shows the percentage of mean square error relative to the total variable values.

The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. As can be observed, the error of the training sets is smaller than that of the validation sets, which is normal because the model was constructed based on the training data. It is interesting to observe that the errors from the validation sets are close to those from the training sets and also represent a small percentage of total variable values. It shows that the models are unbiased which is due to the LASSO method which considers that the bias increases as the shrinkage increases and that the variance decreases as the $\lambda$ increases.

All the models defined by LASSO or by linear regression were validated on a validation set com-

| Mean square error | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARIABLES → | Model size | | DTP (km) | | VIT (km/h) | | $CO_2$ (g/km) | | $NO_x$ (g/km) | |
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 7 | 3,04% | **0,214%** | **0,269%** | 0,337% | 0,180% | 0,227% | 0,318% | 0,222% | 0,250% |
| VIT | 19 | 8,26% | 0,186% | 0,221% | **0,478%** | **0,492%** | 0,221% | 0,307% | 0,209% | 0,274% |
| $CO_2$ | 11 | 4,78% | 0,187% | 0,193% | 0,462% | 0,656% | **0,246%** | **0,371%** | 0,215% | 0,251% |
| $NO_x$ | 11 | 4,78% | 0,186% | 0,193% | 0,423% | 0,625% | 0,215% | 0,324% | **0,233%** | **0,286%** |
| Network mean values | | | *1,31E+04* | *1,31E+04* | *12,77* | *12,81* | *3,23E+09* | *3,21E+09* | *9,69E+06* | *9,68E+06* |

Tab. 3.3 – *The mean squared error of the model established with one variable applied to another.*

pletely different from the training set. However, to be able to compare the results, the training data and the validation data were kept the same throughout this study. The average error remains in the same range with the four sets of links: 2% for $NO_x$ emissions and traveled distance and 3% for $CO_2$ emissions. Therefore, various sampled links or combination between them could provide an estimation of traffic and emission variables over the network with reasonable errors, and allow flexibility in choosing which links to equip.

### Study of the merger and intersection between the network variables

Considering the low sampling rate and the low average error of each variable, and taking into account that they have certain common selected links, it is possible to merge and/or create intersections between sets of link selected for traffic variables and for the pollutant emissions. For example, the selected links identified by the shrinkage method for the traffic data, total traveled distance and spatial mean speed will be joined together (merging of selected links between two variables displayed as $DTP \cup VIT$) to apply a linear regression and obtain a new model with adjusted $\beta$ values for each predictor (links). Merging $CO_2$ and $NO_x$ was considered ($CO2 \cup NOX$) in the same way. Taking into account that some variables have common links, the intersection between them was considered. The advantage of the ($CO2 \cap NOX$) intersection is the possibility of having a model with fewer predictors than the model established by merging. The associated errors were quantified for each resulting variable value. The average error of each linear regression is shown in table 3.4 and was calculated for each variable by considering each situation (merger or intersection).

| Daily average percentage of error (absolute error) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP∪VIT | 26 | 11,30% | 1,92% | 1,78% | 3,01% | 3,48% | 2,24% | 2,55% | 2,23% | 2,18% |
| DTP∩VIT | - | - | - | - | - | - | - | - | - | - |
| $CO_2 \cup NO_x$ | 14 | 6,09% | 1,91% | 1,62% | 5,28% | 5,25% | 2,21% | 2,80% | 2,21% | 2,12% |
| $CO_2 \cap NO_x$ | 8 | 3,48% | 2,00% | 1,68% | 7,00% | 6,61% | 2,41% | 2,98% | 2,38% | 2,10% |
| Number of observations | | | *267* | *133* | *267* | *133* | *267* | *133* | *267* | *133* |

Tab. 3.4 – *The average percentage of absolute error from the linear regression model fitted to the merge or intersection between variables of the same type.*

The links selected for total traveled distance and spatial mean speed are completely different, so

they have no common links. Merging traffic variables allows estimating all the variable values with the same accuracy as the LASSO selection applied to each one separately. The distribution of error values varies from 0.1% to less than 7% considering a confidence interval of 95% in general for all the variables. When the selection in traveled distance shown in table 3.2 is compared with the merger between the traffic variables and with the merger of the pollutant emissions, it can be seen that there are fewer errors. In contrast, the merger models have more links than the traveled distance model fitted by LASSO in table 3.2, which explains the fact that the errors are less dispersed. This appears even more clearly when the mean square errors are compared. The mean square values of the merger and/or the intersection are shown in table 3.5.

| Mean square error | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARIABLES → | Model size | | DTP (km) | | VIT (km/h) | | $CO_2$ (g/km) | | $NO_x$ (g/km) | |
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP∪VIT | 26 | 11,30% | 0,180% | 0,204% | 0,298% | 0,484% | 0,215% | 0,301% | 0,204% | 0,268% |
| DTP∩VIT | - | - | - | - | - | - | - | - | - | - |
| $CO_2 \cup NO_x$ | 14 | 6,09% | 0,185% | 0,193% | 0,415% | 0,625% | 0,214% | 0,324% | 0,208% | 0,265% |
| $CO_2 \cap NO_x$ | 8 | 3,48% | 0,191% | 0,191% | 0,548% | 0,703% | 0,222% | 0,343% | 0,219% | 0,261% |
| Network mean values | | | *1,31E+04* | *1,31E+04* | *12,77* | *12,81* | *3,23E+09* | *3,21E+09* | *9,69E+06* | *9,68E+06* |

Tab. 3.5 – *The mean square error from the linear regression model fitted to the merging or intersection between variables of the same nature.*

When the same comparison is made with the intersection between the selected links of the pollutants, they have almost the same number of selected links and average error values. If we compare the models of the pollutants in table 3.2 with the intersection between them, it is interesting to observe that although it reduces the size of the selection (in this case the selection falls from 11 to 8), the results remain the same.

On the other hand, the merger of the selected links identified by the shrinkage method for the two variables characterizing daily traffic values, and the linear regression model established with them, results in estimating the network daily values with a low average error, using only 11% of the network links. To conclude, by considering only 8 links (3.5% of the network), all the variables can be estimated with an acceptable error: around 2% for traveled distance, 3% for pollutant emissions and less than 7% for spatial mean speed.

### The best model for each variable

Another study was also conducted to know which of all the models studied is the best for each variable, when analyzing their error distributions and model sizes. First, the results are shown for the total traveled distance. Figure 3.10 shows the error distributions of all the models used to calculate the total traveled distances ($DTP$).

Each model is identified on the x-axis, the percentage of errors on the y-axis, and the sizes of the models are shown inside each plot. If a maximal error range of $\pm 5\%$ is considered, three models can be selected: $CO_2$, $NO_x$ and the intersection between the $CO_2$ and $NO_x$ models ($CO_2 \cap NO_x$). The $CO_2$ and $NO_x$ models have the same size: 11 selected links. The $NO_x$ model is less dispersed than the $CO_2$ model. Considering that the aim is to estimate variable values with a minimum number of selected links, the emission intersection model describes the traveled distance better. It has a model

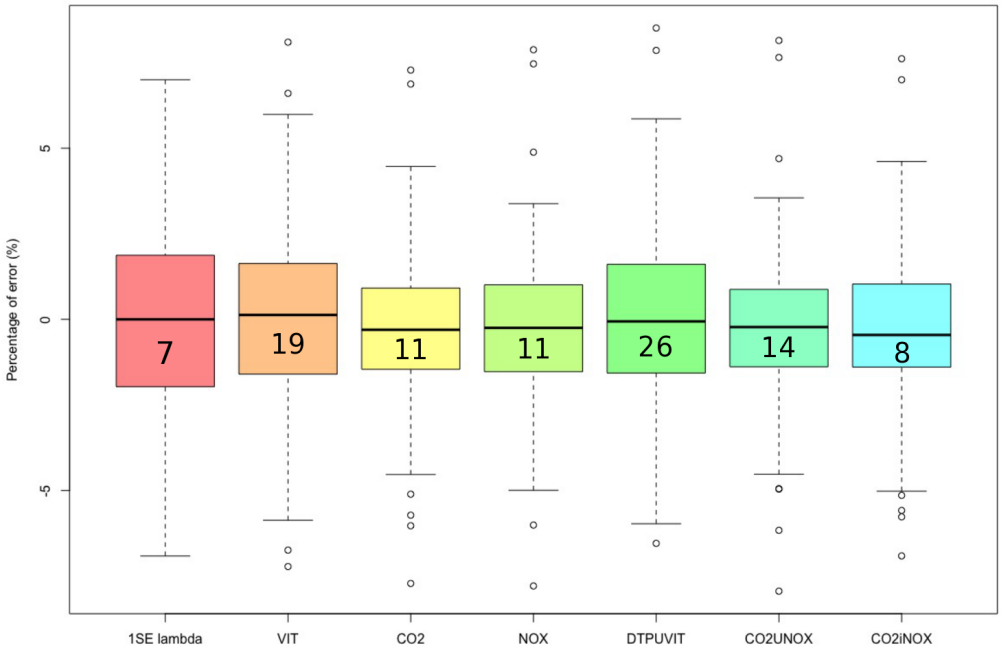



Fig. 3.10 – *The error percentages of model variable selections used to estimate the total traveled distance.*

size of only 8 selected links and its error dispersion is slightly lower than ±5% considering that 50% of the errors are between ±2%.

Figure 3.11 shows the same analysis to determine the spatial mean speed ($VIT$). No model presents an error dispersion between ±5%. The least dispersed model is that built with merged traffic data, thus the selected links from traveled distance and spatial mean speed ($DTP \cup VIT$). This model is composed of 26 links which means 11% of the network is necessary to calculate the spatial mean speed with ±8% error dispersion. The second model capable of describing spatial mean speed is that established by LASSO. This model has 7 variables less than the previous model, but its error dispersion can reach ±10%.

The emissions were also analyzed. As can be observed in figure 3.12, none of the models used to calculate total $CO_2$ emissions has an error distribution between ±5%, which is also the case for spatial mean speed. The model best suited to determining network $CO_2$ emissions is that built from the merger between the traffic variables with an error dispersion of about ±6%. A second possibility is the traveled distance model which has only 7 selected links with an error dispersion of around ±8%. Their error dispersions are similar, but their models sizes are 3 times larger. According to the error range tolerated, it is possible to choose a model that requires traffic information from 7 or 26 links of the network. Considering that the aim is to estimate network variables with less traffic information, the traveled distance model is better suited for determining network $CO_2$ emissions.

Considering network $NO_x$ emissions, only two models have error distributions lower than ±5%: the traveled distance (DTP) and the $NO_x$ models. The model that estimates the $NO_x$ emissions better is the traveled distance model. Compared with the other models, it is the least dispersed and has the smallest number of selected links. This comparison can be seen in figure 3.13.

From all these analyses, it is possible to observe the high correlation between traveled distance

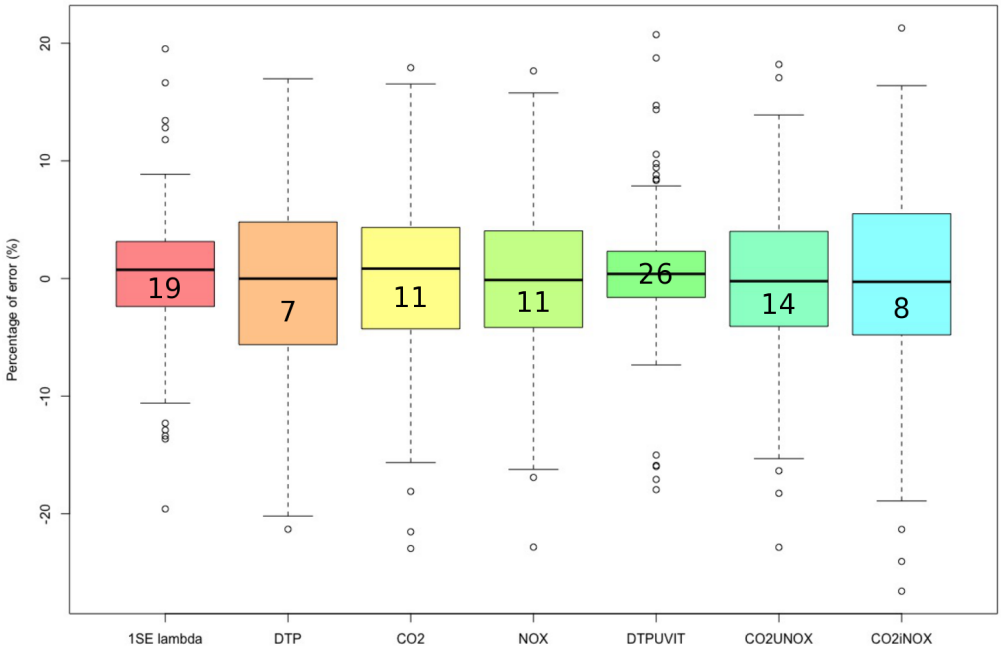



Fig. 3.11 – *Percentage error of model variable selection for spatial mean speed estimation.*

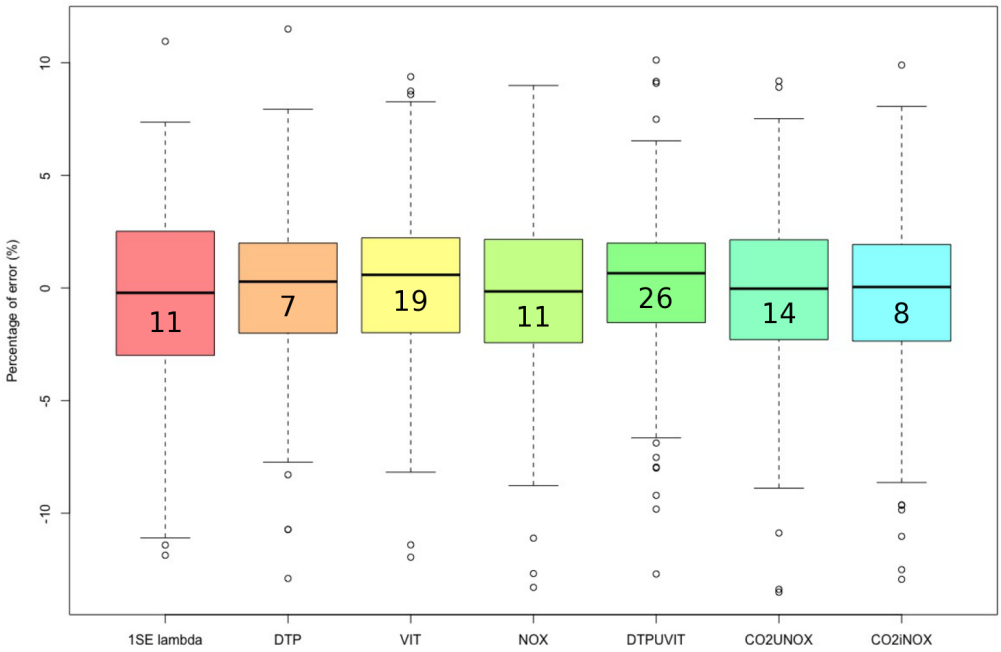



Fig. 3.12 – *The percentage errors of the model selections used to estimate network $CO_2$ emissions.*

and emissions. For both emission variables, the best model was obtained from the selected links in the traveled distance model. When the traffic variables are considered, the $NO_x$ model estimates the network traveled distance better. For spatial mean speed, merging selected traffic data links is the best one. It is possible to conclude that many sets of links can be used to estimate network variables with the help of linear regression performed on the selected links by LASSO. The results are in the

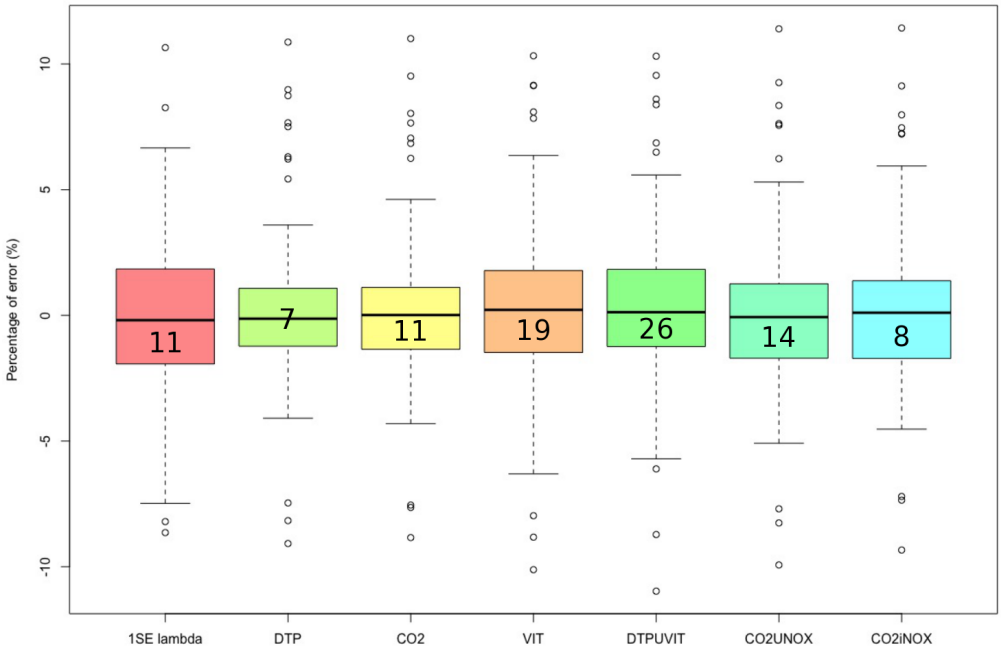



Fig. 3.13 – *The percentage error of the model selections for estimating network $NO_x$ emissions.*

same error range as with LASSO.

A large number of models are studied in these sections and many options were considered to choose the best model for each variable. An index method was used to help choosing between competing models. This model assessment is called the Bayesian information criterion ($BIC$), also known as the Schwarz criterion (Schwarz, 1978). It is defined as follows:

$$BIC = -2 \times L_m + m \times ln(N) \tag{3.7}$$

where $n$ is the sample size, $L_m$ is the maximized log likelihood of the model and $m$ is the number of parameters in the model. The $BIC$ scores measure the degree of accuracy from the fit based on the statistical accuracy of the fit and the number of parameter that have to be estimated (Stanford, 2016). For all the models studied along the static/static dataset, the $BIC$ scores were calculated to compare the performance of all the models together. The lower the $BIC$ score, the better the model (Schwarz, 1978). Table 3.6 shows all the index values for all the models and variables. The scores in red are the lower ones; consequently they are the best models for its respective variables. Considering the traffic variables, the model defined by lambda in the LASSO method in both cases is that which better evaluates the traffic variables, which agrees with our comparison of the densities of the predicted values and the reference ones in figure 3.3 ($a$) and ($c$). For $CO_2$ the model constructed from the merger between selections in both traffic variables gives a better estimation of the emissions than the model that was built based using $CO_2$ pollutant data since the former model had 26 links versus 19 for the latter one (many more selected links).

For the pollutant $NO_x$ we have two possibilities for models, one built with traveled distance data and another with $CO_2$ pollutant data. Both have the same score, so the criterion here will be the model with fewer predictors. In this case it is the traveled distance model with 7 predictors and an error distribution lower than $\pm 5\%$ versus 11 and slightly more dispersed errors than traveled distance

| Bayesian information criterion (BIC) | | | | |
|---|---|---|---|---|
| VARIABLES → | DTP | VIT | CO2 | NOX |
| MODELS ↓ | Validation set | Validation set | Validation set | Validation set |
| Reference | 3925 | 59 | 5469 | 3880 |
| Lasso Lambda | 3721 | -45 | 5327 | 3737 |
| Lasso 1SE | 3739 | -40 | 5341 | 3731 |
| DTP | - | 32 | 5331 | 3726 |
| VIT | 3768 | - | 5322 | 3748 |
| $CO_2$ | 3726 | 39 | - | 3726 |
| $NO_X$ | 3729 | 16 | 5339 | - |
| DTPuVIT | 3748 | -43 | 5317 | 3743 |
| DTP∩VIT | - | - | - | - |
| $CO_2 u NO_X$ | 3729 | 15 | 5339 | 3743 |
| $CO_2 \cap NO_X$ | 3729 | 59 | 5353 | 3737 |
| Min error distribution | 3723 | -39 | 5323 | 3740 |

Tab. 3.6 – *The Bayesian information criterion for each model studied in static/static datasets.*

(DTP).

In general, most of the scores for a given variable are quite similar. This allows choosing models with a good assessment and with fewer selected links. It is possible to conclude that all the models give a good estimation of the variables.

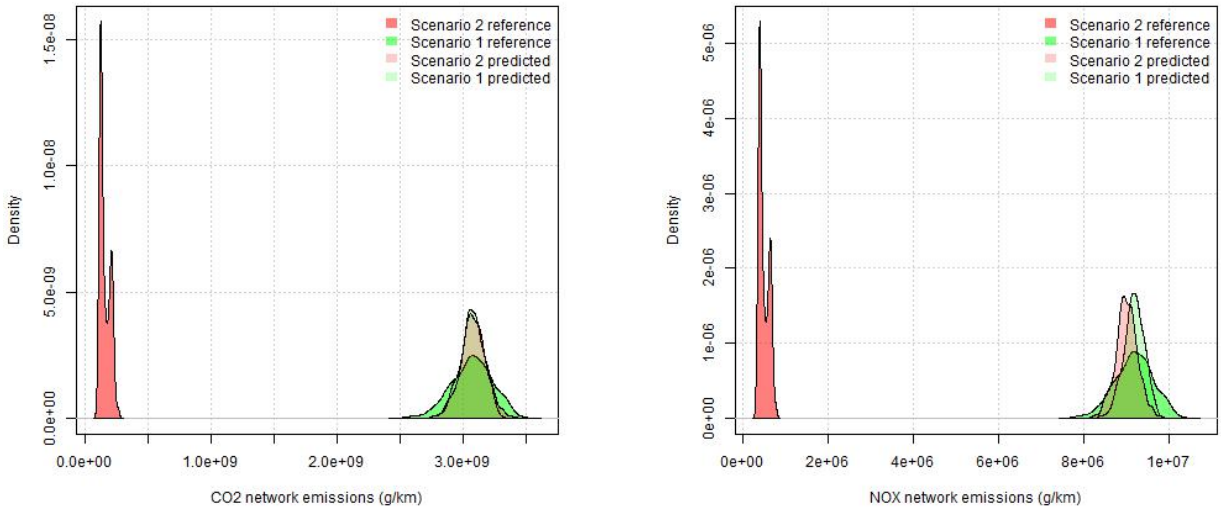
### Influence of route choice on the selection

Each traffic simulation represents the traffic data recovered for each time period (15 minutes). The first 12 time periods represent the peak hours in the morning and the last ones the peak hours in the evening. Both have different patterns for route choices. For more details, the explanations can be found in section 2.1.1 in chapter 2. The effect of the traffic assignment on the selection must be studied. Considering the variables explained previously, the impact of the assignment was studied for each one. Instead of splitting up their observations randomly for each variable into training and validation sets, we used the scenarios to fulfill this role.

Scenario 1 represents the morning peak hours and scenario 2 the evening ones. Two cases were studied: the first uses scenario 1 as the training set and scenario 2 as the validation set; the second case is scenario 2 as the training set and scenario 1 as the validation set. Both scenarios have different route patterns with higher demand in the morning than in the afternoon. The goal is to understand if it is possible to use data from the morning traffic to estimate data in the evening traffic and vice-versa.

Using scenario 1 as the training set and scenario 2 as the validation set and considering the opposite situation, only mean speed converges in this dataset and gives a statistically acceptable result. To understand why linear variables such as traveled distances, and $CO_2$ and $NO_x$ emissions did not converge, the densities of pollutant emissions are shown in figure 3.14 with scenario 1 used as the training set and scenario 2 as the validation set for both cases for traveled distance, and the second case for pollutant emissions. Their respective figures can be found in B.3.

It is possible to observe the considerable difference between morning and evening peak values for both the pollutant emissions and in the traveled distance, when the references are considered. There




(a) $CO_2$ emission densities.

(b) $NO_x$ emissions densities.

Fig. 3.14 – *Pollutant emission densities obtained using the observation of the morning peak as the training set and the evening peak as the validation set.*

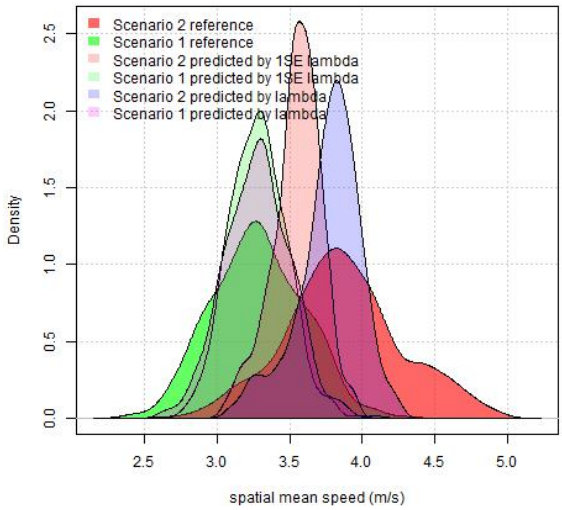
are more emissions in the morning traffic than in the evening. The LASSO method cannot solve this disparity of values. When the method is applied to the morning traffic data and validated using the evening observations, the estimated results are around the median value of the first scenario to which the method was fitted. The same happens in the second case, when fitting the model selection to the evening data and validating for the morning traffic. This is shown for the linear variables in B.3.

Considering the mean speed variable, the result is completely different. The network mean speed does not present such a big disparity as the other variables and has a similar distribution design, though with a slight difference when the median values are considered. Figure 3.15 shows the density values of the network mean speed and their predictions fitted by LASSO and also the error distributions from the predictions.

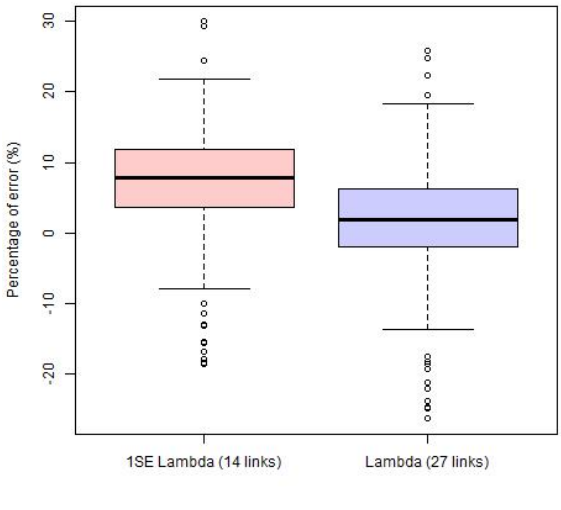
Figure 3.15 ($a$) shows the densities of the reference values of the morning and evening peaks, the values predicted by the model selection fitted in scenario 1 and validated in both scenarios using the two lambda models (regular lambda and 1SE lambda). The morning traffic network presents lower mean speed than in the evening, so the morning traffic was more congested on average. When the evening data was used as a validation set for the model fitted for the morning data, the predicted density using the 1SE lambda model (i.e. model with fewer predictors), showed a distribution between both reference peaks. When the model with more predictors was considered (e.g. lambda model), the distribution of the predicted values was more centered in the median of the reference data of scenario 2. The error distributions of both models are shown in ($b$). The 1SE model selected 14 links of the network considering the morning data. When validated for the evening data, the error can reach from $-10\%$ to $20\%$ with a median around $8\%$. The error distribution of the lambda model, which had almost twice as many selected links compared to the 1SE, was more centered, varying by $\pm 10\%$.

For the second case, when the evening data were used to fit the model that will be applied for the morning data, there were more errors than in the first case. Figure 3.16 shows this case for network mean speed.

Using the evening data as the training set improved the prediction of the network mean speed

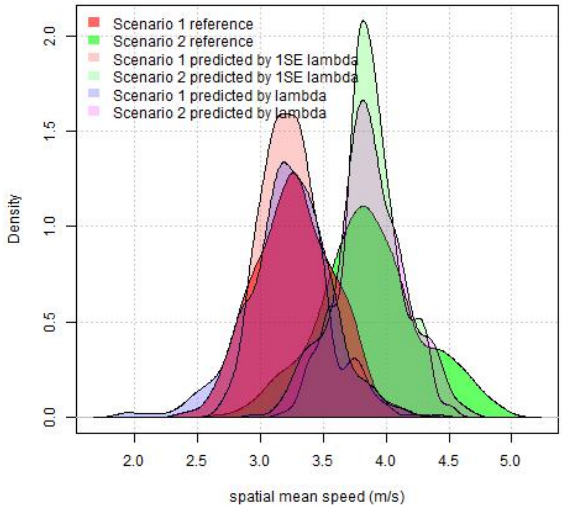



(a) Network mean speed densities for each scenario.

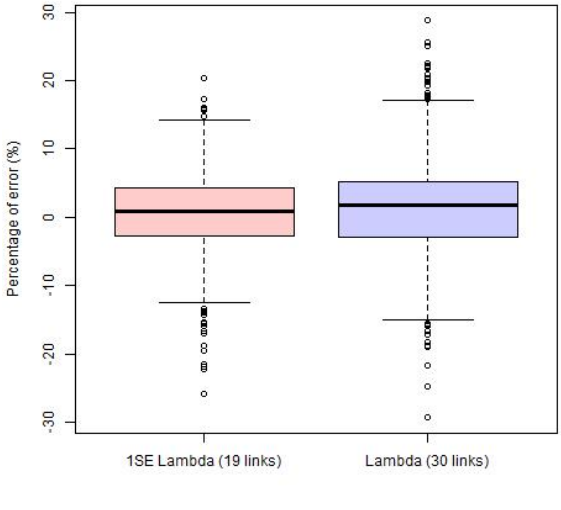



(b) Network mean speed error distributions based on evening data as the validation set.

Fig. 3.15 – *Network mean speed results using scenario 1 (morning) as the training set and scenario 2 (evening) as the validation set.*




(a) Network mean speed densities for each scenario.

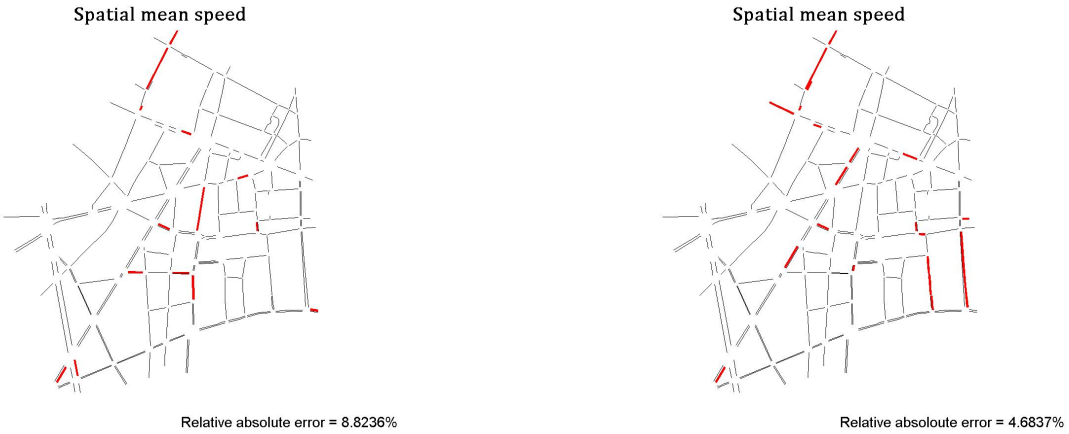



(b) Network mean speed error distributions based on evening data as the validation set.

Fig. 3.16 – *Network mean speed results using scenario 2 (evening) as the training set and scenario 1 (morning) as the validation set.*

for both peak hours more than the first case. The predicted values were centered by the medians of their reference values. The associated errors for both lambda models concerned 50% of the data with errors around ±5%. When considering all the data the error distributions were between ±20%. The second case selected more network links in comparison to the first case. Furthermore, the second case selected the same number of links as LASSO did considering the whole network with both assignment periods aggregated. Considering the 1SE lambda model selection the two cases shared 32% of their

selected links, i.e. 6 links. 14 links were selected in the first case and 19 in the second using the 1SE lambda model. There were 9 links between them (64%) similar to the links selected by the 1 SE model. Together, the morning and evening selection represented 68%, i.e. 13 links out of 19 were identical. The links selected by the 1SE lambda model for both cases are shown in 3.17.




(a) Network mean speed Links selected for scenario 1 (morning).

(b) Network mean speed Links selected for scenario 2.

Fig. 3.17 – *Links selected for mean speed for the morning and evening cases (evening).*

The route choices have a strong influence on the selection method. As observed, for traveled distance and emissions it was not possible to set a unique model for each variable taking into account only morning or evening data to estimate them due to the different traffic dynamics in both cases. Considering the spatial mean speed in the network, it was not significantly different between morning and evening. The morning traffic did not present more dispersed errors than the second case. The evening data estimated the spatial mean speed better for both cases and had fewer errors ±20%, considering that 50% of the data were subject to a percentage errors of about ±5% when estimating morning traffic.

### Conclusion on the static/static dataset

The LASSO shrinkage method was used as a linear regression selection method to select the most relevant links in the network for traffic and emission variables. A model was established for each variable with a set of links with related weightings. Although the variables were very varied over the links and the period of the day, using daily network information as input is sufficient to estimate the variables accurately. The analysis concluded that the links selected for traveled distance present better results in terms of model size when estimating daily traffic and emission values.

In the light of these considerations, two conclusions can be reached: (*i*) the strong correlation between traveled distance and spatial mean speed allows determining both pollutant emissions on the basis of their samples, because they are dependent on both traffic variables; and (*ii*) the fact that both can be used to determine other variable values using a simple linear regression, leads us to conclude that there is no single acceptable sample (set of links). To sum up, we can deduce that it is not necessary to rely on the finest traffic data to quantify and determine daily traffic and emission values at network level, since the daily values are sufficient to do this. In our study, the sampling method provided a model that uses ±5%, on average, of the network's equipped links to obtain the traffic data, and it can estimate variables with errors within ±5% and ±10%.

This showed that links could be interchanged without compromising the estimation, demonstrating flexibility in the selection of the links used to estimate network values in this case study. The model with the fewest selected links will be the best choice, especially from the practical point of view, for transportation managers when they decide to implement sensors at network links.

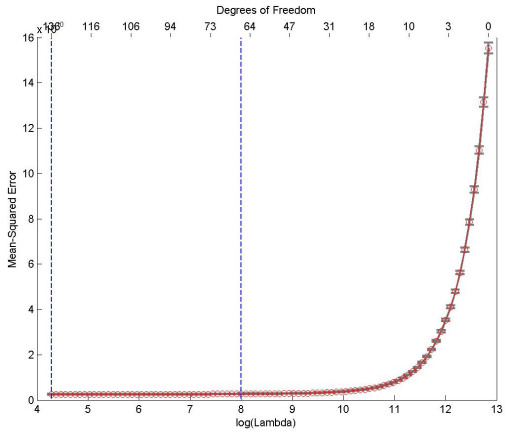
### 3.3.2 LASSO applied to static/dynamic datasets

#### Models proposed by LASSO

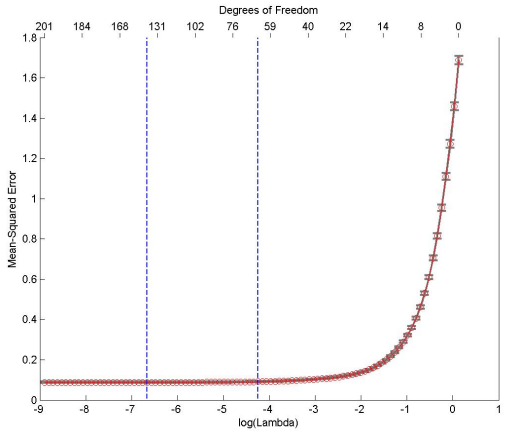
As explained before, the LASSO selection highlights the two possible models capable of providing the optimal compromise between accuracy-few errors and sparsity-lower size selection. The first model is called lambda (associated with the minimum error) and the second is called 1SE lambda (with fewer selected links). The comparison between them will be made for each variable in this dataset.

This dataset has 4 different variables: traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions. These variables are structured as $n \times p$ matrix, where $p$ are the links represented in the network while $n$ are the 15-minute observation values for each link. Each link has 9600 observations (400 observations for each period of time; each link has 24 periods of time) and they are split up randomly into two parts: the first represents 2/3 of the matrix and it was used as a training set to apply LASSO; the second part, 1/3 of the original matrix, represents the validation set where the LASSO selection will be validated and the errors associated will be quantified.

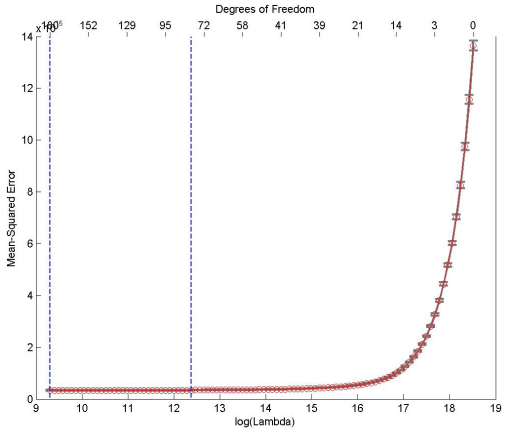
The mean-square prediction error curve for traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions are shown below.



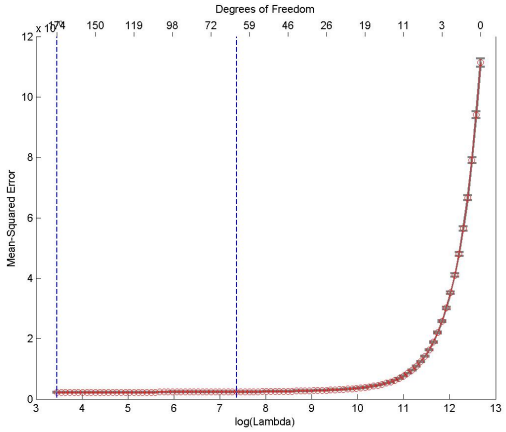

(a) Traveled distance mean-square prediction error curve




(b) Spatial mean speed mean-square prediction error curve




(c) $CO_2$ emissions mean-square prediction error curve




(d) $NO_x$ emissions mean-square prediction error curve

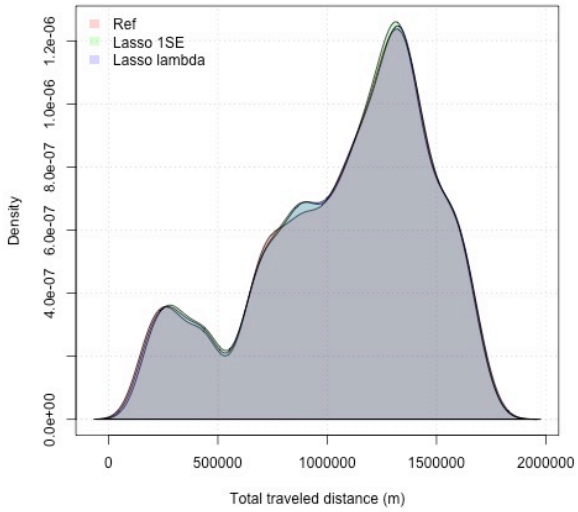
Fig. 3.18 – *Estimated prediction error curves and their standard errors for the variables in static/dynamic datasets.*

As can be seen in figure 3.18 each curve is plotted as a function of the corresponding ($\lambda$) value. The horizontal axis describes the model's complexity which increases as we move from right to left. The estimates of prediction errors and their standard errors were obtained by tenfold cross-validation, *i.e.* the training sets were divided into ten equal parts to apply the cross-validation method. The optimal models are indicated by the vertical lines on the plot. The top of each plot is annotated with the size of models.

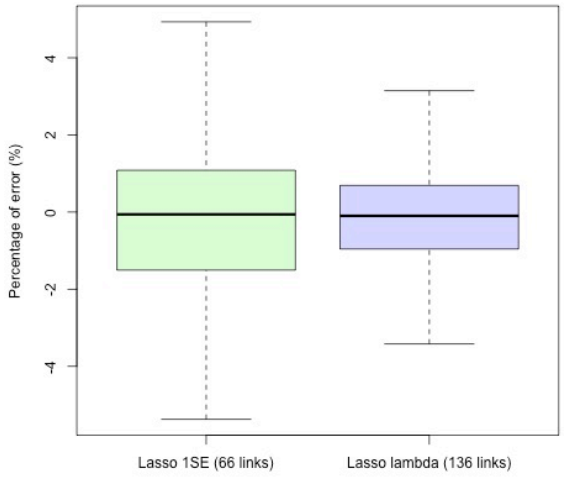
Indeed, for our purposes, it is better to have a model with a minimum number of predictors (links). The relative errors of the resulting values of each model will be compared to confirm that this choice can be made.

The resulting values of all the variables (traveled distances, spatial mean speeds and emissions) of the model with lambda choice with minimum error (the left line in 3.18) and the one standard-error model (the rightmost vertical line in 3.18) have the same rate of data explanation, about 98%, indicating the good precision of the model.

The predicted values for each variable in each lambda model were compared with their respective reference values and the associated errors were calculated. Figure 3.19 shows the densities and associated errors of the model in comparison to the reference values for each variable.

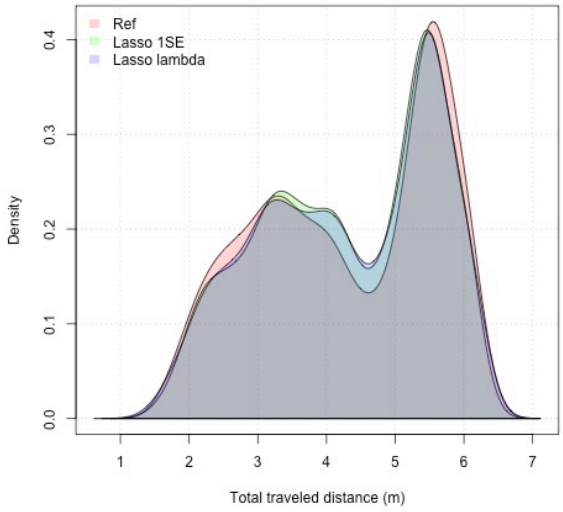



(a) Traveled distance densities.

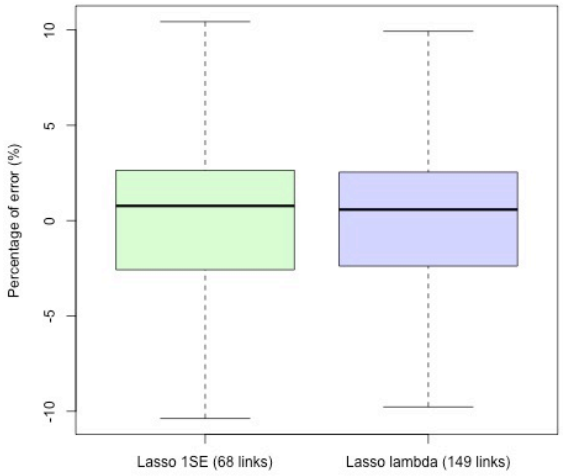



(b) Traveled distance error distribution by lambda models.

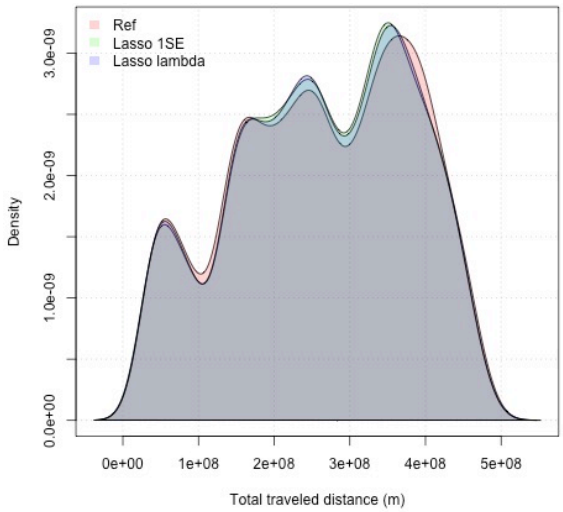
The densities of the models fitted by LASSO show a good prediction for all the variables but mainly for traveled distance. The error distributions for both the models proposed are quite similar for each variable. They all present some outliers and are shown in C.1. Considering these outliers the errors can reach from $-100\%$ to $50\%$. In figure 3.19 $(b)$, $(d)$, $(f)$, $(h)$ show the same error distributions but without outliers. The traveled distance is related to an percentage error lower than $\pm 5\%$, the spatial mean speed and $CO_2$ emissions around $\pm 10\%$ and $NO_x$ lower than $\pm 10\%$. The lambda model is almost 3 times the size of the 1SE lambda model with a similar error range. The 1SE lambda model selects, on average, 30% of the network versus 69% for lambda model. As with the previous dataset, the 1SE lambda model was selected to continue the study for this dataset as its size is smaller while its error range is similar.



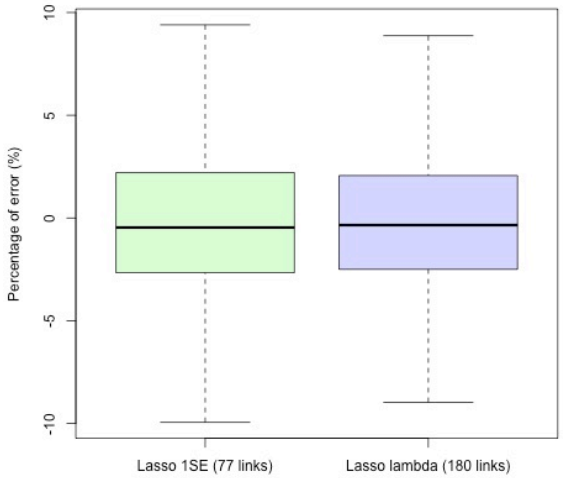

(c) Spatial mean speed densities.




(d) Spatial mean speed error distribution by lambda models.




(e) $CO_2$ emission densities.



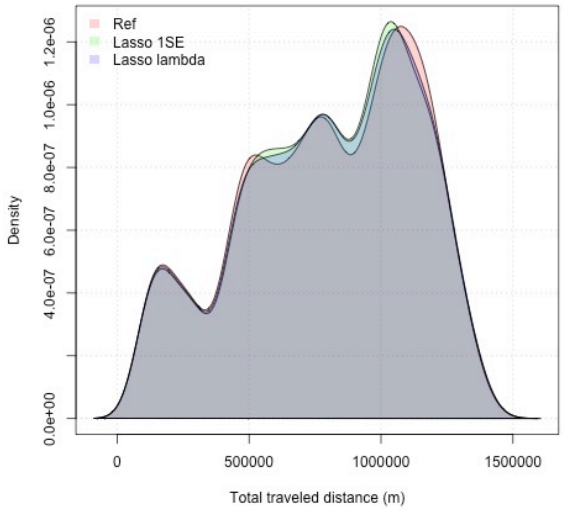

(f) $CO_2$ emission error distribution by lambda models.

### The 1SE model

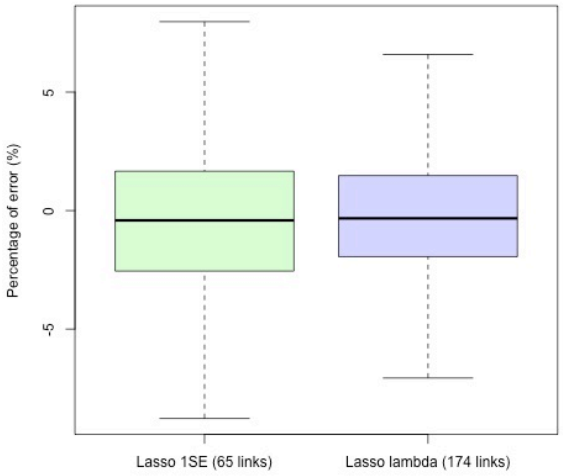
As explained before, the model proposed by $\lambda$ with one standard error from the minimum square error was the model chosen for all the variables. The results are presented only for this model.

LASSO selected 66 links (out of 230) for the traveled distance. This explains 98% of the data used for observation considering the confidence interval of 95%; 68 links for spatial mean speed considering 95% of the data explained by the model; 77 links for $CO_2$ emissions with a model that explains 98% of the data; and $NO_x$ emissions with 65 links selected to define a model explaining 98% of the data. The selected links on each variable are displayed in figure 3.20 and their average absolute percentage error given by the model in figure 3.21.

Considering the links selected for all the variables, 14 links are identical. If we compare the Links selected for traveled distance with those selected for both pollutant emissions, there are 28 common

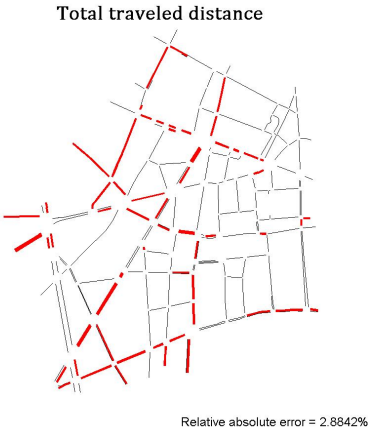



(g) $NO_X$ emission densities.

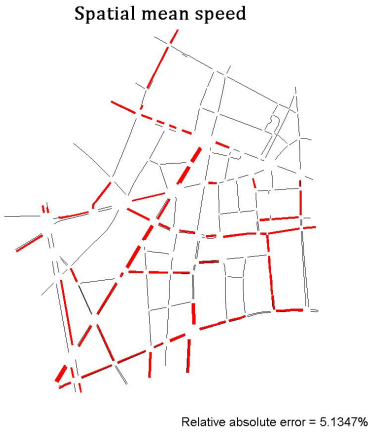



(h) $NO_X$ emission error distribution by lambda models.

Fig. 3.19 – *Comparison between the reference values and the predicted values and their associated errors.*




(a) Selected traveled distance links

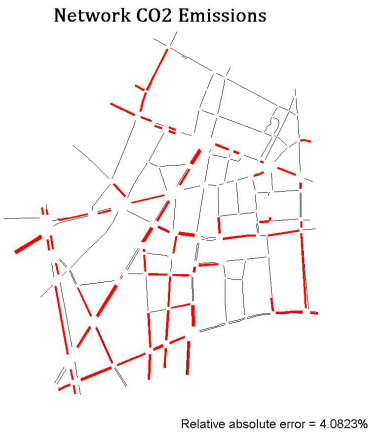



(b) Selected spatial mean Speed links

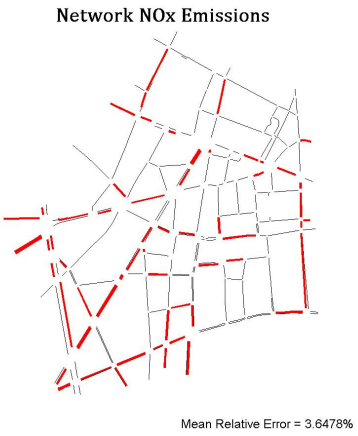
links. Similarly, if we compare the selection of spatial mean speed with both pollutant emission selections, the three variables have 23 common links. Table 3.7 gives the percentage of common links selected between the variables.

| VARIABLES → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| MODELS ↓ | Validation set | Validation set | Validation set | Validation set |
| DTP | 100% | 45,5% | 54,5% | 57,6% |
| VIT | 44,1% | 100% | 50,0% | 42,6% |
| $CO_2$ | 46,8% | 44,2% | 100% | 59,7% |
| $NO_x$ | 58,5% | 44,6% | 70,8% | 100% |

Tab. 3.7 – *Ratio of common selected links between variables.*




(c) $CO_2$ selected emission links




(d) $NO_x$ selected emission links

Fig. 3.20 – *Selected links determined by the λ model calculated with one standard error rule and the absolute average percentage of errors given by the model using the selected links.*

It is interesting to observe that out of the 77 Links selected for $CO_2$ emissions, only 15 are not the same for any of the other variables and the rest are a combination between links selected for spatial mean speed and traveled distance. The same occurs for the $NO_x$ selection, only 6/65 links are selected that are completely different in comparison to the other variables. If we compare the selection between both pollutants, it is possible to observe that there are at least 20 different links between them. This allows concluding that each pollutant must be analyzed separately and that we cannot consider a single set of links to quantify both pollutants even if they have 70% links in common.

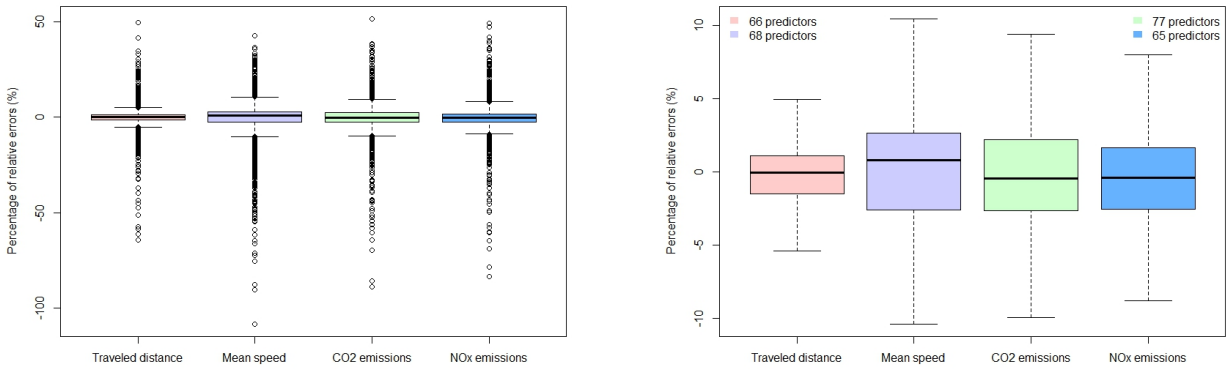
We also compare the links selected with the most contributive links presented in chapter 2 for each variable. For traveled distance, of the 66 links selected, 13 are part of the most traveled links on the network. Similarly, for the spatial mean speed variable, 27 out of 68 links have an average value equal to or lower than 10km/h; the $CO_2$ selection has 15/77 links among the most contributive. The result is 13/65 for $NO_x$. As with the static/static datasets, this comparison leads us to conclude that the selection is not concentrated only on the most traveled or pollutant links at the network.

The structure of the static/dynamic dataset allows estimating the total network emissions for 15-minute periods from the selected links. The network variable values are calculated using the model with the selected links fitted to a training set applied to the validation set. Their results were compared and the associated errors calculated. They are shown in figure 3.21.

The errors of the linear variables such as traveled distance and pollutant emissions were centered around zero compared to the spatial mean speed which presented a right-skewed distribution with a median around 1%. Without their outliers, the least dispersed error distribution is traveled distance with errors around ±5%. Then the spatial mean speed and pollutant emissions are around ±10%. Considering their outliers, the error distribution reached between -100% and 50%. These outliers represented a several traffic situations over the 3200 estimated, but more than 85% of the traffic and emission had errors within the normal range distribution.

### Model size and error distribution

Considering that all the models studied in this dataset selected many more links than the static/static dataset, new lambda models were studied for each variable. The goals were: (*i*) to reduce the number of selected links in the models; (*ii*) identify models that had an error distribution between ±5%;




(a) Percentage error of all variables considering their outliers

(b) Percentage error of all variables without outliers

Fig. 3.21 – *Percentage error between the predicted variable values and the response values.*

and (*iii*) identify an optimal model that had fewer links than the models studied previously with a reasonable error distribution.

The traveled distance model has 66 links considering the one-standard error of lambda (*) and 136 links considering the lambda model (**). Figure 3.22 shows the error distribution of the LASSO lambda models that have fewer predictors than the 1SE model. The models for all the variables with their outliers will be presented in appendix C.2 and their conclusions will be drawn in this section.

As can be seen, the error distributions are less scattered when the size of the model increases. Considering the outliers for the same models presented in 3.22 from the objective function (0) to 6 links, the outliers appear only in the negative error range. For model sizes with more predictors, the outliers appear on both sides of the box plot, and for models with more than 21 predictors selected, the error distributions seem stable until the 2 star model. Of all models displayed, the least dispersed is the model fitted with lambda (**). Figure 3.23 shows the models which have selected links between the 1SE and lambda models.

The model fitted to the 1SElambda model (*) had errors between -60% and slightly more than 40%. If the outliers are not taken into account, around 85% of the data have errors of ±5%. The model with 66 predictors shows 50% of its data with errors of less than ±2%, while from this model until the lambda model (**) the error distributions considering the outliers are the same. When only the box plots are considered, the difference is that the error distribution of the lambda model is around ±4% instead of a range between ±5% for the 1 SE lambda and it has twice as many predictors as the 1 SE lambda model (*).

For the spatial mean speed, the error distributions are more scattered than the traveled distance. The errors, considering the outliers, can easily reach more than 100% error as shown in figure C.3. When analyzing the error distribution of the same models without taking into account the outlier values, the same trend observed in the models for traveled distance does not occur, indicating a linear reduction of the error dispersion as the links increase inside the models. Each model presents a singular distribution of errors, as can be seen in figure 3.24. The models with 56, 68 and 149 links have similar error distributions.

As explained in the beginning of this section, the models between the two optimal lambdas defined by LASSO were investigated. Figure 3.25 shows the error distribution of these models. None of the

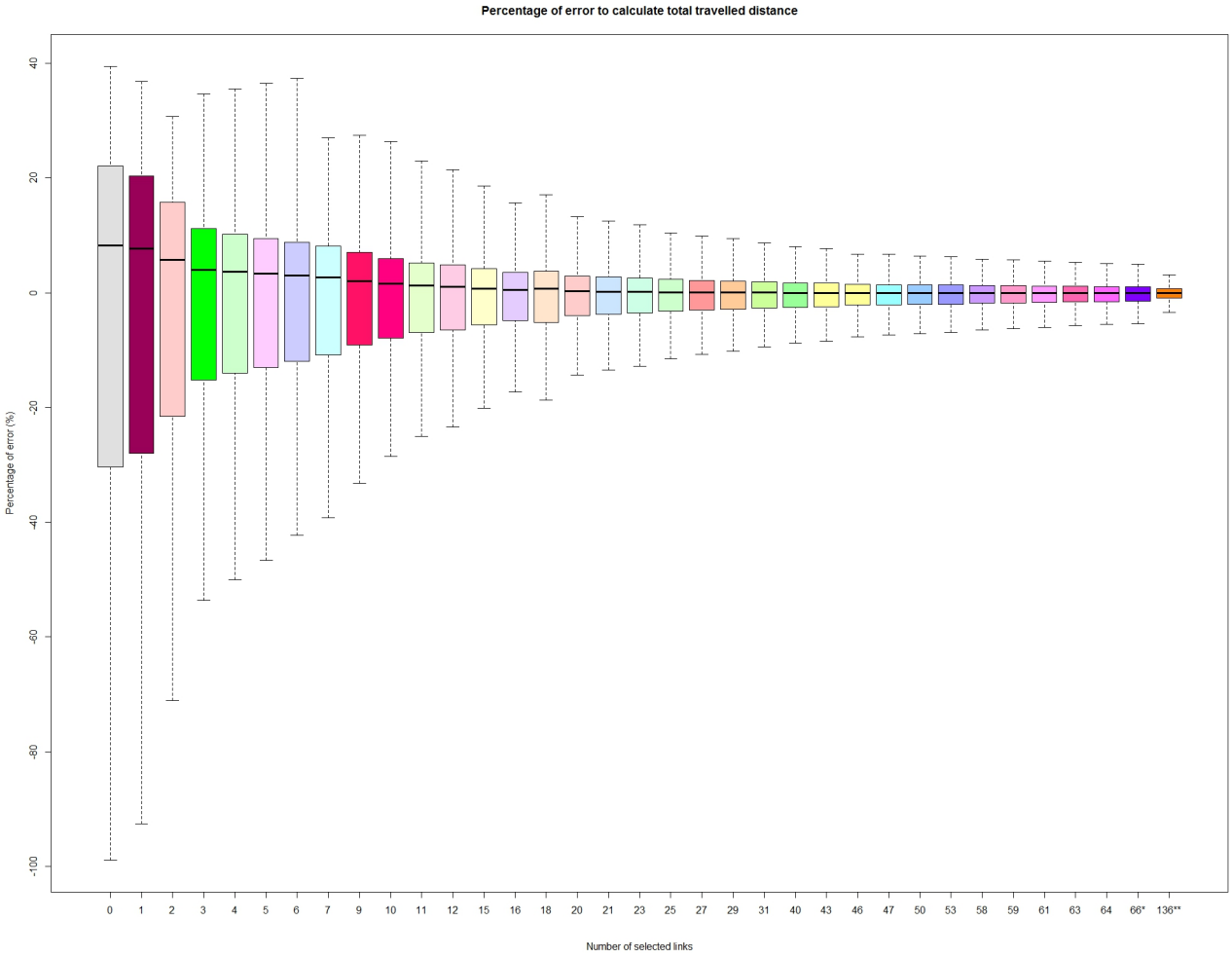
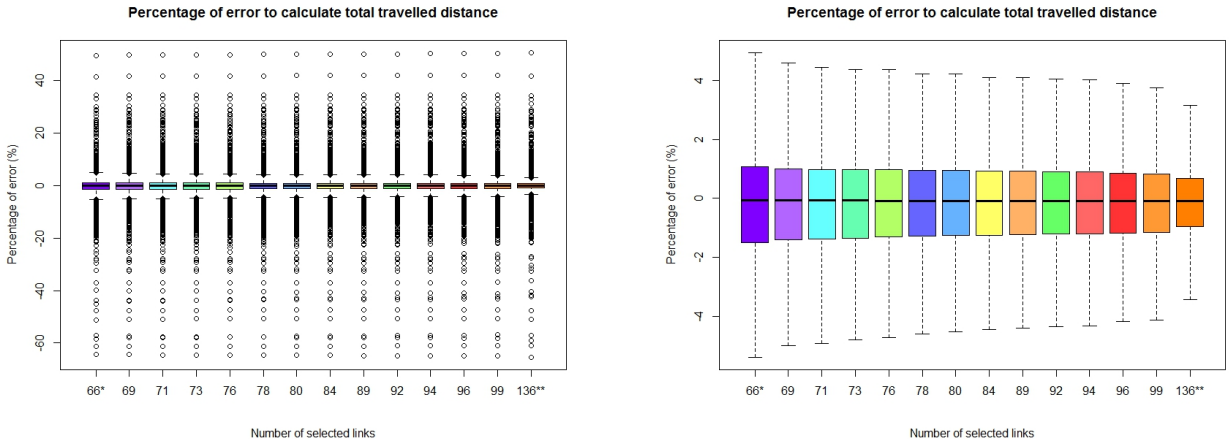



Fig. 3.22 – *The traveled distance error distributions from lambda models fitted using LASSO and that have fewer predictors than the 1 SE lambda model (all plots without their respective outliers).*




(a) Error distributions of models between the two possibilities of lambda determined by LASSO.

(b) Error distributions between both optimized lambda models without their outliers.

Fig. 3.23 – *Error distributions on the traveled distance models.*

models have errors between ±5% and all of them have almost the same distribution when outliers are considered. The lambda model (**) contains 149 predictors, which represents 65% of the network. For this model, 15% of the data are outliers and have error distributions of about ±10%, considering 85%

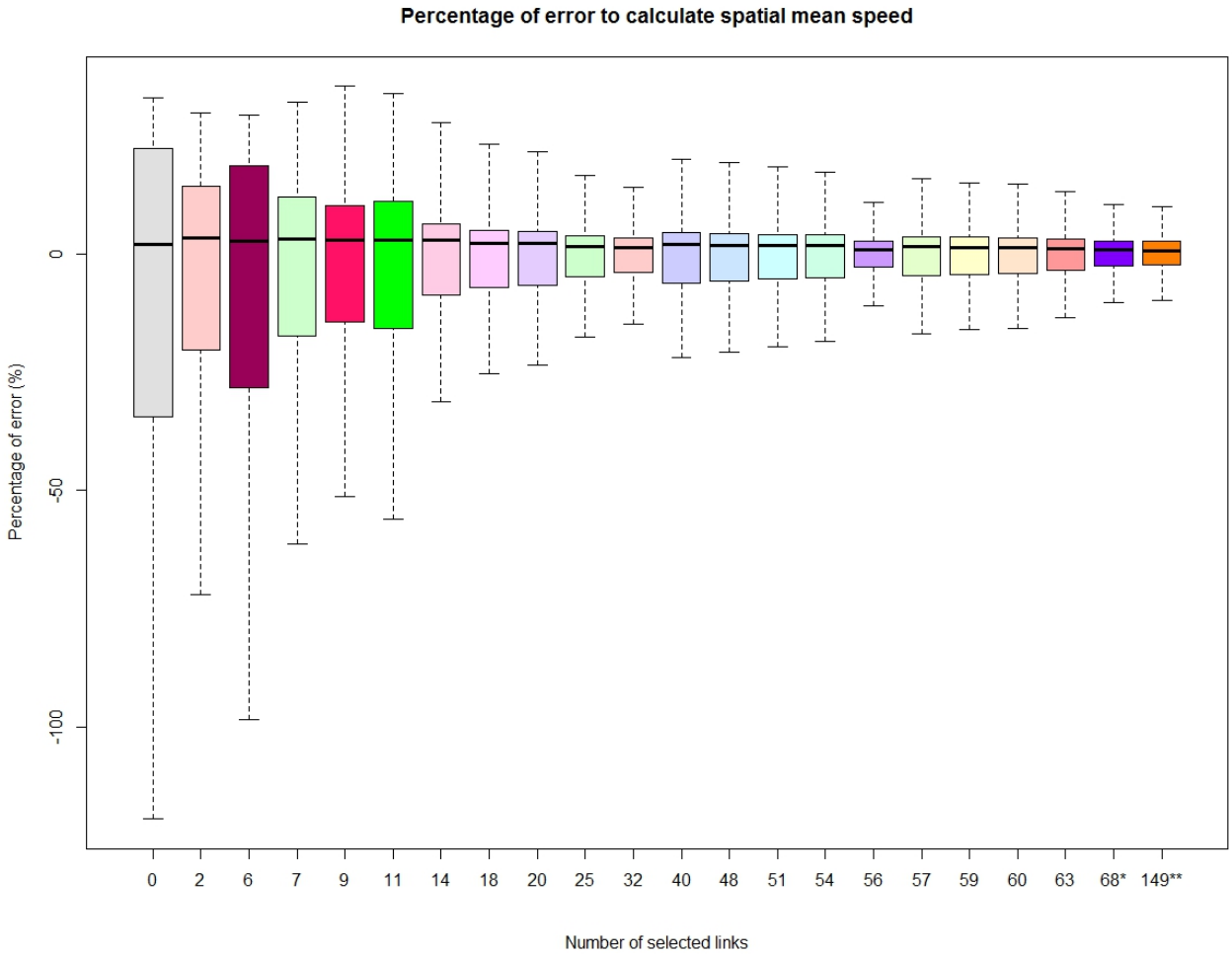
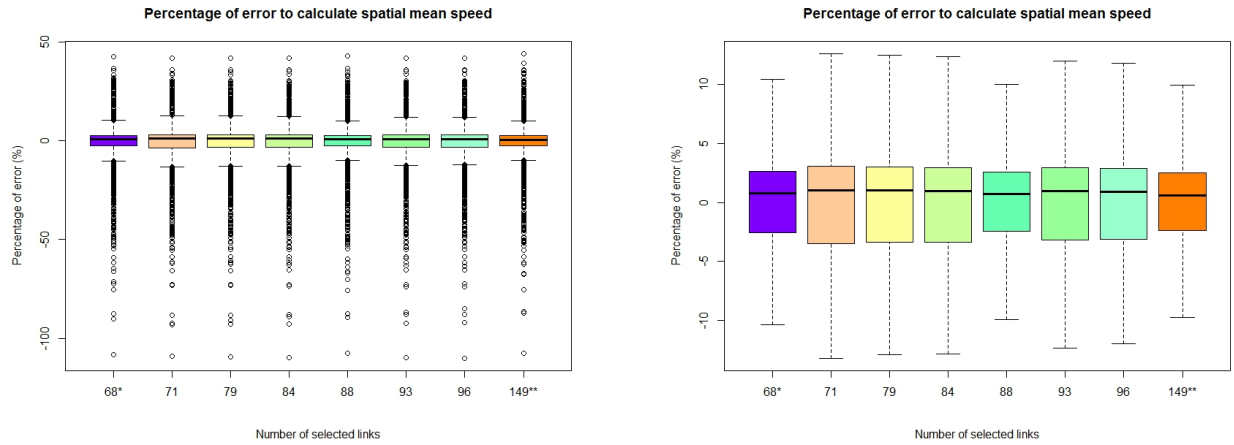



Fig. 3.24 – *The error distributions of spatial mean speed for all lambdas with models with fewer predictors than 1SE lambda.*

of the remaining data. It is also interesting to observe that all the models from zero to 149 predictors tend to underestimate the spatial mean speed even if the outliers are considered.




(a) Error distributions of models between the two optimal lambda models determined by LASSO.

(b) Error distributions between both optimized lambda models without their outliers.

Fig. 3.25 – *Error distributions on the spatial mean speed models.*

Analysis of the pollutant emissions in this dataset shows that the $CO_2$ selection has more Links

selected for LASSO optimized lambdas than the other variables; 77 links for the 1 SE lambda model and 180 links (78% of the network) for the lambda model. Figure C.4 shows the models which have fewer selected links than defined by the one-standard error lambda model. These models follow the same trend as traveled distance models, indicating that the larger number of predictors in the model results in a less scattered error distribution. The models from that with 19 selected links to the lambda model have similar distributions when outliers are considered, as shown in C.4. The objective function reached errors as high as 800% only for the $CO_2$ variable, and taking the outliers into account. Figure 3.26 shows the same figure but without outliers. Regarding the model containing 43 predictors, all of them have almost centered distribution and are less scattered than for the previous models.

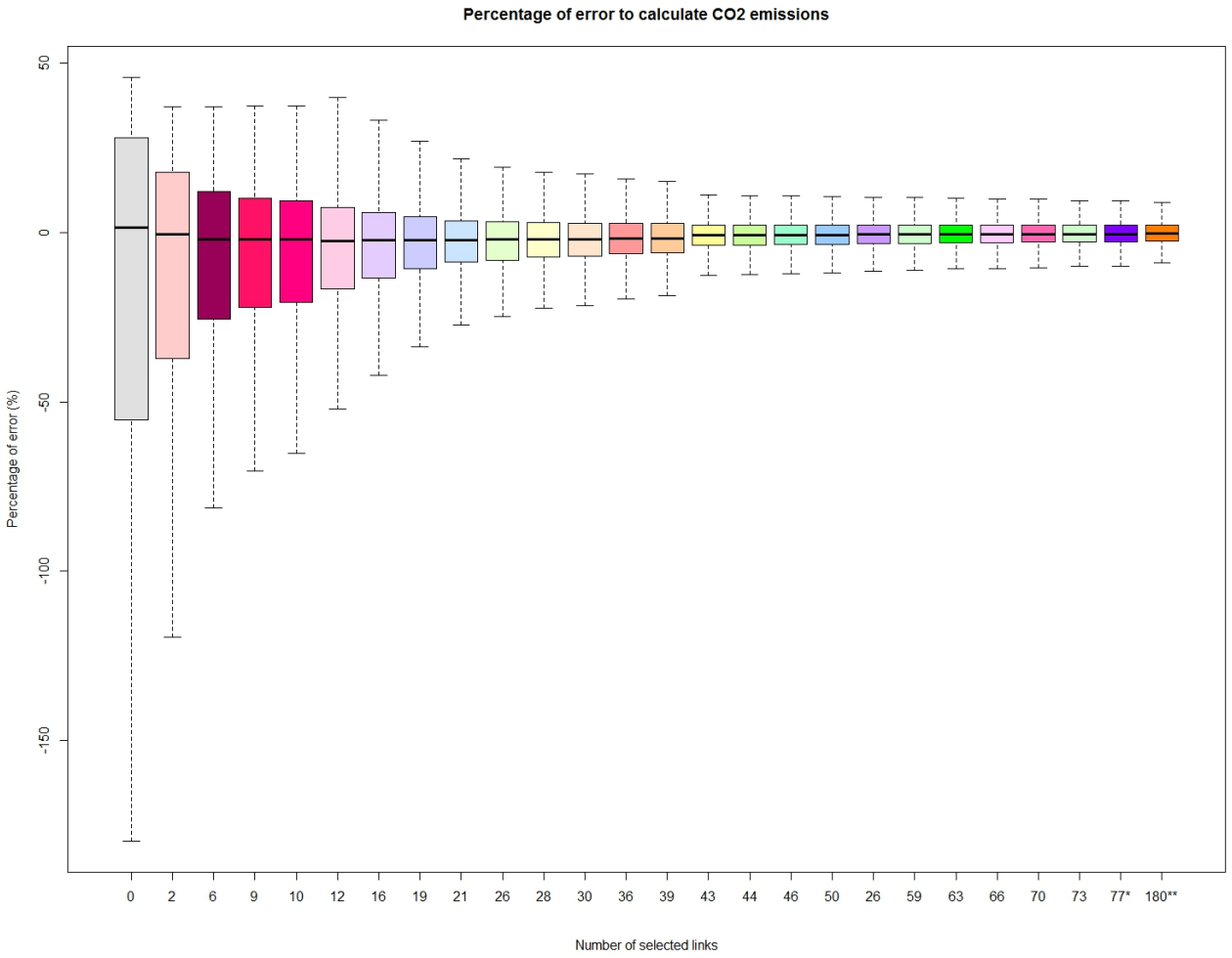
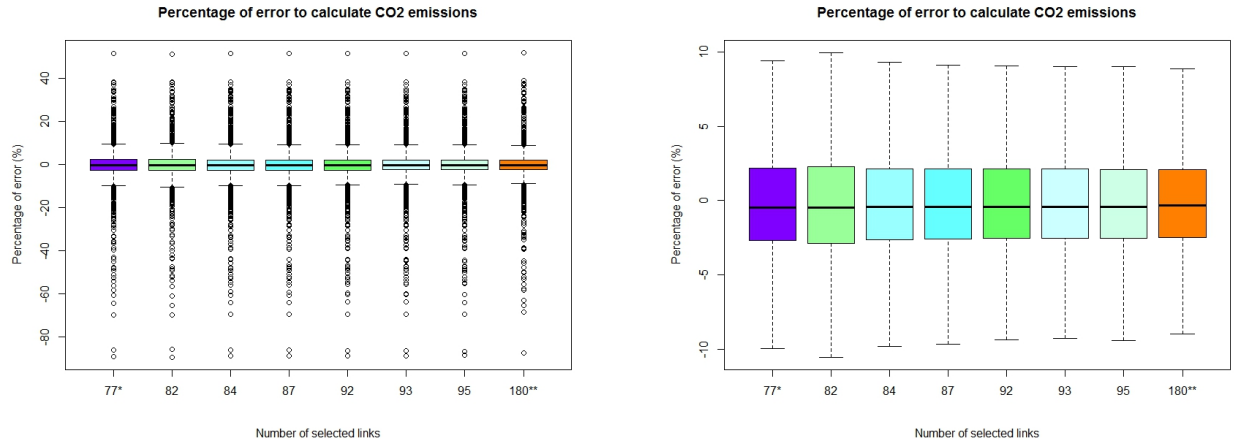



Fig. 3.26 – *The error distributions from network $CO_2$ emission models fitted using LASSO and that have fewer predictors than the 1 SE lambda without outliers.*

Considering the models between both optimized lambdas defined by the LASSO method at the beginning of this chapter (in figure 3.18), all of them have almost the same error distribution despite the huge difference in model size. None of them have errors distributed between -5% and 5% even when considering 180 links in the model. All these observations are represented in figure 3.27.

The same trend analyzed in the $CO_2$ also applied to the $NO_x$ models. The LASSO defined two optimized models to estimate 15 minute emissions for the network: the 1 SE lambda with 65 predictors inside the model and the lambda model with 174.

The models with fewer predictors than the 1SE lambda model are shown in figure C.5 and 3.28 without outliers. They also tend to underestimate emissions in the models with between 0 and 39




(a) Error distributions from models between the two possibilities of lambda determined by LASSO.

(b) Error distributions from models between the two possibilities of lambda determined by LASSO without their outliers.

Fig. 3.27 – *Error distributions on the network $CO_2$ emission models.*

selected predictors and their error distributions are more centered.

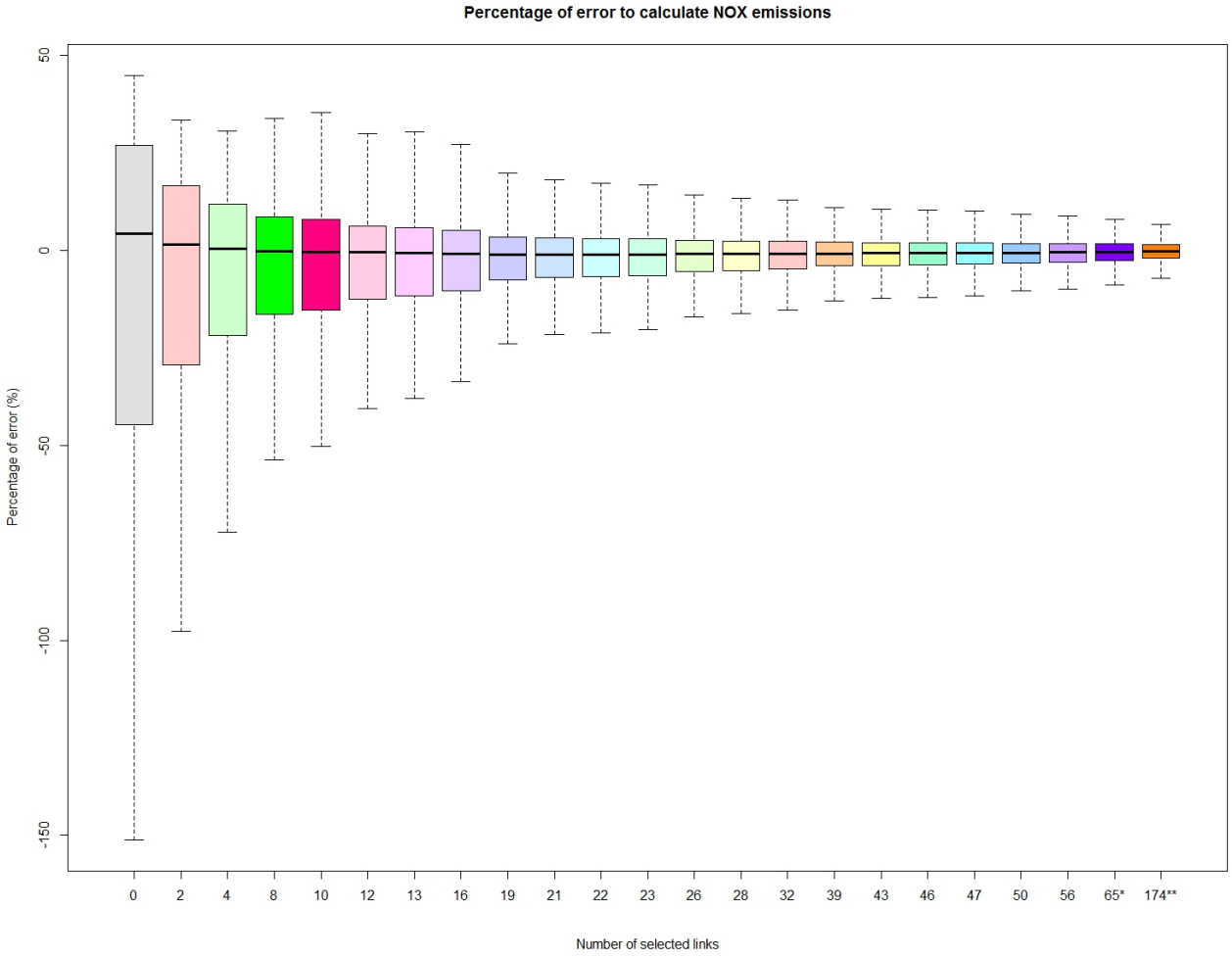
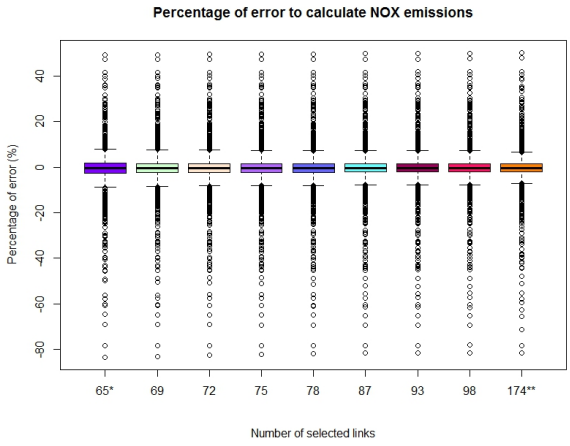


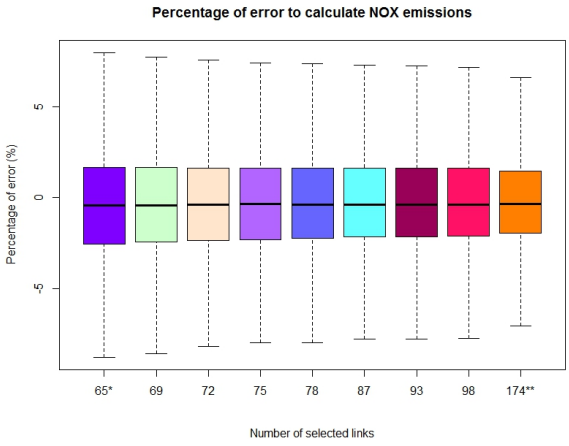

Fig. 3.28 – *The error distributions from network $NO_x$ emission models fitted using LASSO and that have fewer predictors than the 1 SE lambda without outliers.*

Considering the models between both optimal lambdas, all of them have outliers that can reach

percentage errors between -80% and slightly more than 40%. Taking into account only the boxplots, they have error distributions of about ±5%. The differences between both optimized lambda models are: the lambda model has almost three times as many selected links as the 1SE lambda model and the errors are slightly less dispersed. These models are shown in figure 3.29.




(a) Error distributions from models between the two optimal lambdas determined by LASSO considering the outliers.




(b) Error distributions from models between the two optimal lambdas determined by LASSO without their outliers.

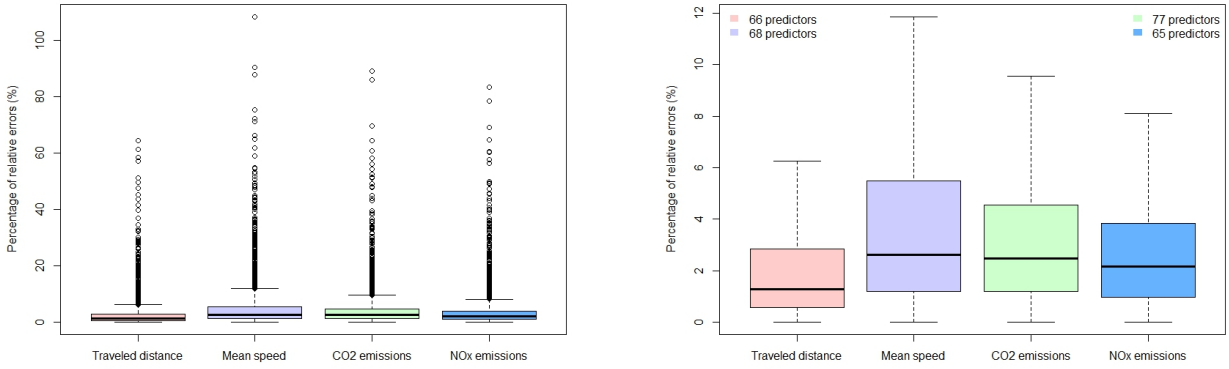
Fig. 3.29 – *Error distributions on the network $NO_x$ emission models.*

Considering model size versus error distribution leads to the conclusion that: (i) the traveled distance and both pollutant emissions have the same trend, which means that a larger model size reduces the width of the error distribution; (ii) for all the variables, the least dispersed model is the lambda (**), but it requires more than 60% of the network links to estimate network variables; (iii) the errors of the 1SE lambda model are slightly more dispersed but this model has less than half the number of predictors of the lambda model. All the variables presented models with 30 selected links with reasonable error.

### The robustness of the models

Considering the high number of observations for each link (3200 in the validation set), and also taking into account the high variability of values representing the periods of free flow and congestion for each link, it is normal to observe more outliers for each variable. This high variability is due to the fact that a single observation can represent a free-flow or a congested state for the same link. For example, a very short traveled distance in the link for period 2 and a high value of traveled distance on the same link in period 7. To respond to this variability, LASSO selects many more links compared to the previous dataset to allow estimating the variable values inside the confidence bounds. The traveled distance required 28.7% of links to build a model with an absolute average percentage error of 2.88%. More than 50% of the data have a percentage error lower than 3% and their outliers represent 15.2% of the validation set. 29.6% of the network links of the spatial mean speed model are selected, which represents an absolute average error of about 5%. Considering all the observations, the percentage error of the model obtained is lower than 6% error for more than 50% of the observations, and its outliers represent 12% of the validation data. For the pollutants, $CO_2$ emissions have a model size with 77 links and an absolute average error around 4.1%. Considering the error distribution, more than 50% have errors below 4.5% and the outliers represent 7.3% of the observations. In contrast to $CO_2$

emissions, the $NO_x$ model needs fewer predictors, the model size contains 65 links with an absolute mean error of around 3.7% and more than 50% of the data have less than 4% error. The outliers from the $NO_x$ model represent 7.9% of the validation set. Even considering the high variability of the 15-minute data values and all the models built using lambda and 1SE lambda, 97% of the data values can be obtained within the confidence interval with an absolute mean error at about 4%. The static/dynamic datasets are better fitted than the static/static datasets. For the models fitted to the static/static datasets, only 75% of data can be obtained considering the confidence interval with an absolute average error lower than 5.50%. All these conclusions are shown in figure 3.30.




(a) Error distributions from the absolute error values considering the outliers.

(b) Error distributions from the absolute error values without considering the outliers from the 1SElambda model.

Fig. 3.30 – *Error distributions from absolute error values of the 1SE lambda model.*

A cross-analysis between variables was conducted to observe the possibility of a single set of selected links to estimate other variable values using a linear regression. Table 3.8 shows the average absolute error between the models established for one variable and then applied to others.

| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 66 | 28,70% | 2,91% | 2,88% | 5,43% | 5,49% | 4,12% | 4,06% | 3,35% | 3,42% |
| VIT | 68 | 29,57% | 2,40% | 2,40% | 5,07% | 5,13% | 4,07% | 4,04% | 3,40% | 3,45% |
| $CO_2$ | 77 | 33,48% | 2,34% | 2,34% | 5,08% | 5,19% | 4,13% | 4,08% | 3,25% | 3,28% |
| $NO_x$ | 65 | 28,26% | 2,36% | 2,35% | 5,50% | 5,49% | 4,11% | 4,08% | 3,58% | 3,65% |
| Number of observations | | | 6400 | 3200 | 6400 | 3200 | 6400 | 3200 | 6400 | 3200 |

Tab. 3.8 – *Percentage error of a model applied to a variable.*

As explained previously, the LASSO method was applied to the training set, which represents 2/3 of the dataset, the resulting model was applied to the validation set and then the absolute average errors were calculated. In table 3.8 both average errors are shown. The red values represent the mean errors of the LASSO model. The other ones represent the average absolute errors corresponding to the model fitted for the variables defined by the links and then applied to the variables defined in the columns. In general, their absolute errors are in the same range as the LASSO model. Even the validation set had similar error values compared to the respective training sets to which the LASSO was fitted. A

square error was calculated to show the difference between the training and the validation set clearly. Both square error values (training and validation) are similar, which proves the good performance of the model. It is also shown that the strong correlation values between all the variables allows any set of selected links to estimate other network values. The mean square values are shown in C.1.

### Study of the merger and intersection between network variables

A combination between the selected links of two variables was considered. The merger and intersection between traffic variables and pollutants were taken into account. The intersection of all the variables is intended to considerably reduce the number of selected links inside the models. The mean absolute errors of this study are shown in table 3.9.

| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP∪VIT | 104 | 45,22% | 2,35% | 2,35% | 4,91% | 5,07% | 4,00% | 4,01% | 3,29% | 3,35% |
| DTP∩VIT | 30 | 13,04% | 3,01% | 3,00% | 5,64% | 5,65% | 4,63% | 4,62% | 3,94% | 4,00% |
| $CO_2 \cup NO_x$ | 95 | 41,30% | 2,34% | 2,34% | 5,02% | 5,15% | 3,94% | 3,95% | 3,23% | 3,30% |
| $CO_2 \cap NO_x$ | 47 | 20,43% | 2,38% | 2,36% | 5,58% | 5,56% | 4,18% | 4,13% | 3,36% | 3,40% |
| DTP∩VIT∩$CO_2$∩$NO_x$ | 13 | 5,65% | 4,02% | 4,12% | 6,54% | 6,50% | 5,45% | 5,47% | 4,90% | 4,94% |
| Number of observations | | | *6400* | *3200* | *6400* | *3200* | *6400* | *3200* | *6400* | *3200* |

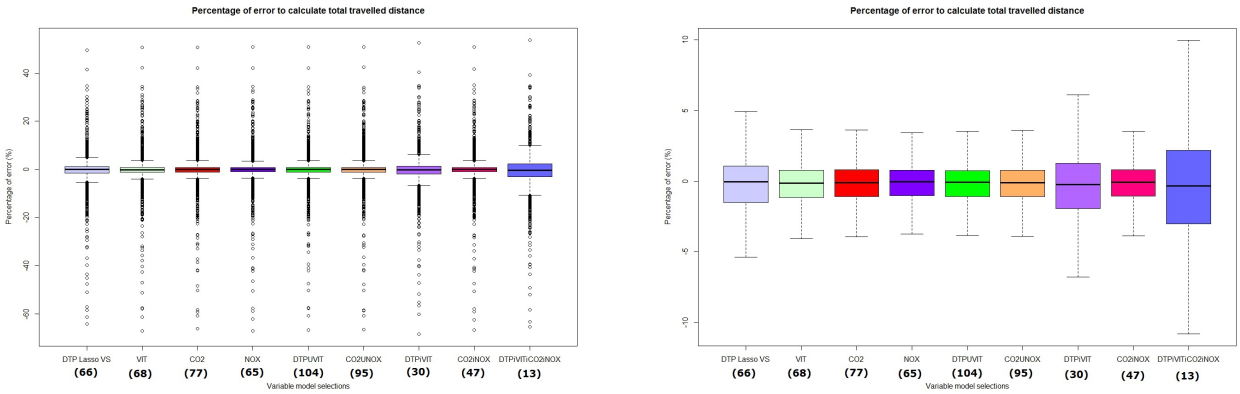
Tab. 3.9 – *Percentage of absolute error from a model applied in a variable.*

The merger between traffic variables and between pollutants increases the size of the model and it can include more than 40% of the network links. Their mean absolute error values are similar to those of the 1SE LASSO lambda model. When the intersection is considered, it is possible to reduce the model size by 13% for the traffic variable intersection and by 20% for the pollutants. Both have absolute errors in the same range as the 1SE LASSO lambda model for the respective variables. To further reduce the number of predictors in the models, an intersection between all the variables was considered. With only 13 links which represent 5.7% of the network, the absolute errors were slightly higher (about 1%) than those of the LASSO model. Therefore, considering only the links corresponding to the intersection may be sufficient to estimate the global variables while significantly reducing the model's size.

### The best model for each variable

All the models were compared to determine the best one for estimating each single variable, i.e. traveled distance, spatial mean speed and pollutant emissions. The first variable considered was traveled distance. Figure 3.31 shows the error distributions of each model estimating the network traveled distance over a time range of 15 minutes.

When all the models are compared with their outliers, they all present almost the same error distribution between 40% and -60%. If the confidence bounds are taken into account, it can be seen that the error distribution corresponds to more than 80% of the data. For traveled distance, the $NO_x$ model has the smallest error distribution in comparison to all the other models. More than 50% of the data have errors between -1.50% and lower than 1.50%. In addition, the model is based on 65 links, just one predictor fewer than the 1SE lambda model. If an error distribution between -5% and 5% is considered, the pollutants intersection model ($CO_2 \cap NO_x$) can be chosen to estimate the traveled




(a) Percentage error of all traveled distance models considering their outliers.

(b) Percentage error of all traveled distance models without outliers.

Fig. 3.31 – *Error distribution of models used to determine the network traveled distance.*

distance using 47 links instead of 65. Depending on the error range, the traveled distance can be estimated using one of these models, even that which takes into account only 13 links with an error range between -10% and 10% ($DTP \cap VIT \cap CO_2 \cap NO_x$).

For the spatial mean speed, considering the outliers in each model, the errors can reach more than -100% and 50%, and they are more dispersed than the traveled distance models. These error distributions are shown in figure 3.32 without considering their outliers and in figure C.6 considering the outliers in the error distributions.

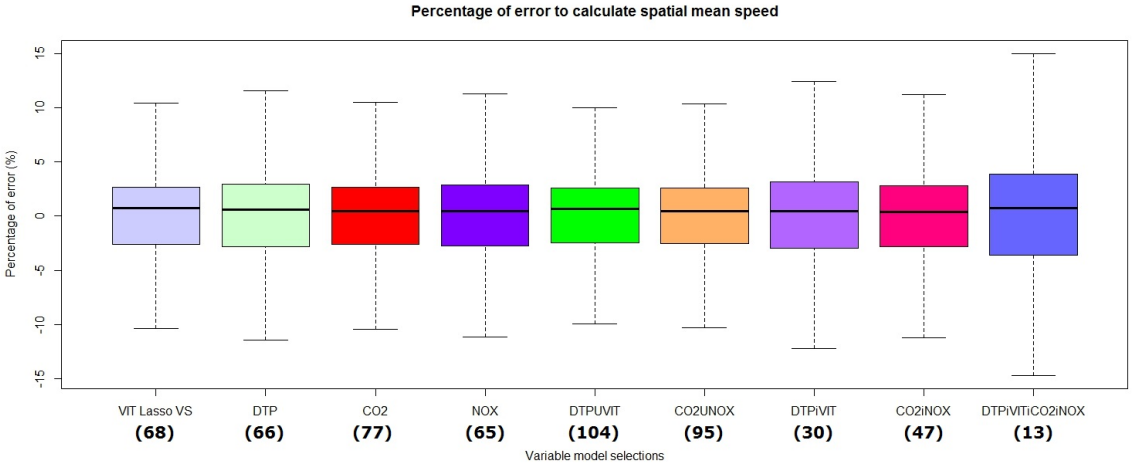



Fig. 3.32 – *Error distribution of models that estimate the spatial mean speed without outliers.*

When the error distributions are analyzed without their outliers, they are all similar. The least scattered is the model built by merging the traffic variables, with errors between -10% and slightly lower than 10%. This model considers 45% of the network links, more than the other models studied. None of these models have an error distribution lower than ±10%.

Considering all the models, more than 80% of their data is distributed similarly. $CO_2 \cap NO_x$ can be considered as a good choice of model, as it comprises a good compromise between model size and error distribution, and a percentage error of slightly more than ±10%. Depending on the level of error tolerance when estimating the spatial mean speed, one option is the intersection of all the ($DTP \cap VIT \cap CO_2 \cap NO_x$) variables since this model has only 13 links with a percentage error of ±15%. All these conclusions can be seen in figure 3.32.

All the models used to estimate $CO_2$ had similar error distributions. When the outliers are removed, it can be seen that all the boxplots have a similar error distribution with different model sizes (number of links). Only the $DTP \cap VIT$ (intersection between traffic variables) and $DTP \cap VIT \cap CO_2 \cap NO_x$ models have slightly higher error distributions than the others. This can be explained by their size, as they have half as many links as the other models. None of them have a percentage error lower than ±5%, and they are all in the range of ±10%. These conclusions can be seen in figures C.7 and 3.33.

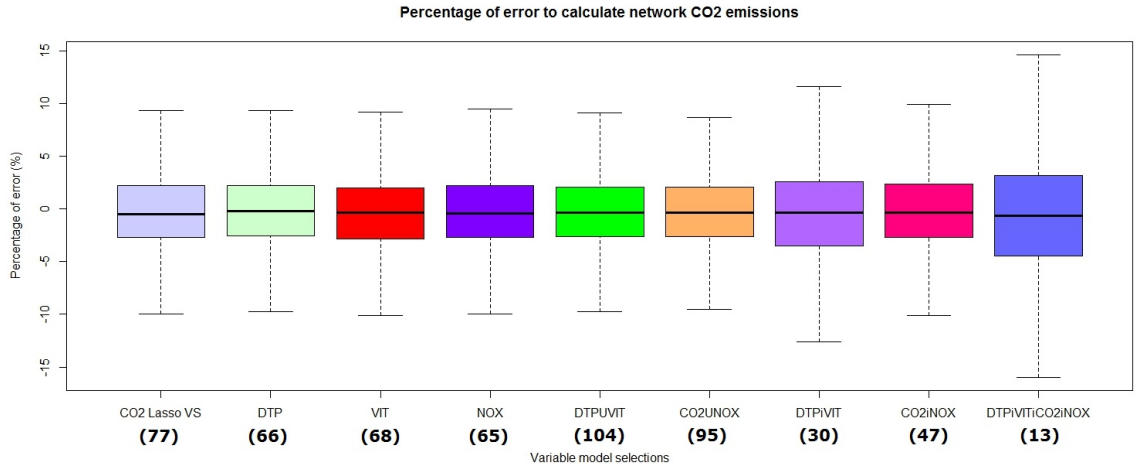



Fig. 3.33 – *Error distributions of models used to estimate the network $CO_2$ emissions without their outliers.*

The same results can be observed for $NO_x$ emissions: (i) the same error distributions considering the outliers with errors between -80% and 40%; (ii) they also present a similar distribution without their outliers in the same figure; (iii) $CO_2 \cap NO_x$ (intersection between pollutants) is the model that best describes the network $NO_x$ emissions, with 47 links, fewer than the 1SE lambda model fitted to the $NO_x$ data; (iv) none of them have errors around ±5%. It is interesting to observe that the results of most of the models are better than the model fitted by LASSO. This leads us to conclude that there is more than one possible set of links to estimate the variable values. These conclusions can be seen in figure C.8 and the error distributions without outliers in 3.34.

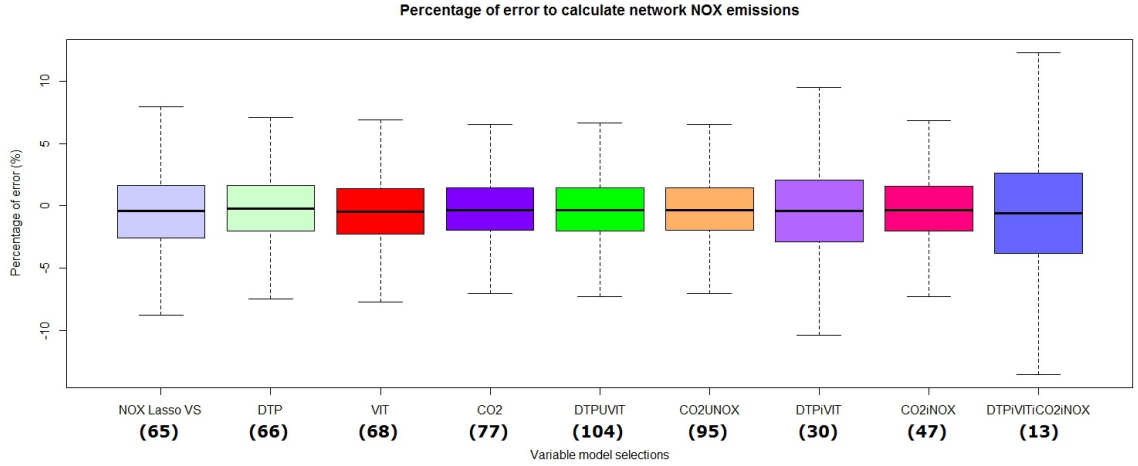



Fig. 3.34 – *Error distribution of models used to estimate the network $NO_x$ emissions without their outliers.*

All the possible models presented in the figures have outliers. They were analyzed to determine to what extent they are present in the validation set. Table 3.10 shows the percentage of outliers for each model studied. These percentages are based on the 3200 observations.

| VARIABLES → | Model size | | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Validation set | Validation set | Validation set | Validation set |
| Lasso training set | - | - | 15,58% | 12,97% | 7,41% | 7,94% |
| Lasso validation set | - | - | 15,22% | 11,97% | 7,31% | 7,88% |
| DTP | 66 | 28,70% | - | 11,84% | 7,69% | 9,03% |
| VIT | 68 | 29,57% | 14,31% | - | 7,16% | 8,69% |
| $CO_2$ | 77 | 33,48% | 13,97% | 12,31% | - | 9,25% |
| $NO_x$ | 65 | 28,26% | 15,69% | 12,50% | 7,53% | - |
| DTPuVIT | 104 | 45,22% | 14,91% | 12,53% | 7,50% | 8,88% |
| DTP∩VIT | 30 | 13,04% | 10,44% | 10,97% | 6,00% | 6,88% |
| $CO_2$u$NO_x$ | 95 | 41,30% | 14,31% | 12,47% | 7,59% | 9,28% |
| $CO_2$∩$NO_x$ | 47 | 20,43% | 15,06% | 12,97% | 7,06% | 9,38% |
| DTP∩VIT∩$CO_2$∩$NO_x$ | 13 | 5,65% | 8,09% | 10,25% | 4,59% | 6,38% |

Tab. 3.10 – *Percentage of outliers in each model.*

On average, all the models except $DTP \cap VIT$ and $(DTP \cap VIT \cap CO_2 \cap NO_x)$, have between 10% and 12% outliers in their validation set. In comparison, the proportion of outliers in the other models is between 7% and 9%. If other variables are considered, the traveled distance showed an average of 14% outliers, all models taken into account, followed by the spatial mean speed with 12%, then $NO_x$ and $CO_2$ with 8% and 7% respectively. In conclusion, the models with intersections of all the variables had fewer outliers on average than the others and the model with $CO_2$ data was that with the fewest outliers.

In table 3.11, the average errors of the training set and the validation set clearly show the difference between them. As expected, the training sets contain fewer errors than the validation sets. Here, it is interesting to observe the extent to which these values represent the network mean values. In addition, both the training and validation values are quite similar, knowing that the values of the data in the validation sets are different from those of the training sets. This means that, on average, all the models were well fitted. The red values correspond to the model fitted by LASSO.

According to the number of variables and the error distribution threshold we defined, a large choice of possible models and sets of links can be used to estimate a variable. In general, all the models fitted have similar error dispersions without considering their outliers. Because of the considerable variations affecting the data (describing traffic and emission data every 15 minutes in periods of free flow and congestion), they contain a minimum of 10% outliers, depending on the model or variable.

It is not sufficient to focus on the errors to compare the models. The complexity, *i.e.* the number of parameters, is also a very important factor. Consequently, we evaluated all the models using $BIC$ criteria. Table 3.12 shows the $BIC$ values for each model.

Compared to the reference values, the lowest scores highlighted the best model (in red). For the traveled distance variable (DTP) the LASSO model defined by lambda defines the variable better than the others. This model contains 136 selected links and has an error distribution lower than ±4%. This

| VARIABLES → | Model size | | DTP (km) | | VIT (km/h) | | $CO_2$ (kg) | | $NO_x$ (kg) | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 66 | 28,70% | 0,060% | 0,086% | 0,090% | 0,129% | 0,088% | 0,122% | 0,076% | 0,107% |
| VIT | 68 | 29,57% | 0,059% | 0,085% | 0,084% | 0,129% | 0,087% | 0,119% | 0,075% | 0,107% |
| $CO_2$ | 77 | 33,48% | 0,058% | 0,084% | 0,064% | 0,129% | 0,088% | 0,120% | 0,074% | 0,105% |
| $NO_x$ | 65 | 28,26% | 0,058% | 0,084% | 0,084% | 0,129% | 0,088% | 0,121% | 0,077% | 0,108% |
| DTP∪VIT | 104 | 45,22% | 0,058% | 0,084% | 0,084% | 0,129% | 0,086% | 0,119% | 0,074% | 0,106% |
| DTP∩VIT | 30 | 13,04% | 0,062% | 0,089% | 0,090% | 0,129% | 0,093% | 0,129% | 0,080% | 0,113% |
| $CO_2 \cup NO_x$ | 95 | 41,30% | 0,058% | 0,084% | 0,084% | 0,122% | 0,085% | 0,119% | 0,074% | 0,106% |
| $CO_2 \cap NO_x$ | 47 | 20,43% | 0,058% | 0,084% | 0,090% | 0,129% | 0,088% | 0,121% | 0,075% | 0,107% |
| DTP∩VIT∩$CO_2$∩$NO_x$ | 13 | 5,65% | 0,070% | 0,102% | 0,103% | 0,141% | 0,102% | 0,141% | 0,089% | 0,125% |
| Network mean values | | | *1,07E+03* | *1,08E+03* | *15,52* | *15,56* | *2,61E+05* | *2,63E+05* | *7,85E+02* | *7,97E+02* |

Tab. 3.11 – *Mean square error of each model and variable.*

| VARIABLES → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| MODELS ↓ | Validation set | Validation set | Validation set | Validation set |
| Reference | 91598 | 10772 | 128111 | 90488 |
| Lasso Lambda | 78468 | 1407 | 115837 | 78035 |
| Lasso 1SE | 78649 | 1509 | 115976 | 78197 |
| DTP | - | 1651 | 116039 | 78111 |
| VIT | 78530 | - | 115905 | 78096 |
| $CO_2$ | 78481 | 1522 | - | 78027 |
| $NO_x$ | 78480 | 1691 | 115987 | - |
| DTP∪VIT | 78481 | 1445 | 115900 | 78057 |
| DTP∩VIT | 78876 | 1742 | 116421 | 78483 |
| $CO_2 \cup NO_x$ | 78479 | 1499 | 115912 | 78031 |
| $CO_2 \cap NO_x$ | 78499 | 1731 | 155994 | 78119 |
| DTP∩VIT∩CO2∩NOX | 79744 | 2281 | 117007 | 79096 |

Tab. 3.12 – *The Bayesian information criterion for all models in static/dynamic dataset.*

is followed by three models that have similar scores: $CO_2 \cup NO_x$, then $NO_x$ and $DTP; \cup VIT$. The first has 95 links inside the model with almost the same error distribution as the lambda model. The second, the $NO_x$ model, has 65 predictors (1 predictor fewer than defined by the 1 SE lambda model) and with error distribution bounds similar to the previous ones. The $CO_2$ model and $DTP \cup VIT$ had the same score, considering 77 and 104 links inside the model respectively. Both have an error distribution of $\pm 4\%$.

Considering the spatial mean speed variable, the lambda model has 149 selected links and a percentage error slightly lower than $\pm 10\%$ in the model with the best Bayesian score. The scores of the other models are much higher than the lambda model. The same occurs with the $CO_2$ variable: the model with 180 predictors and an error distribution of $\pm 10\%$ (lambda model) has the lowest score compared to the others. None of them have similar scores.

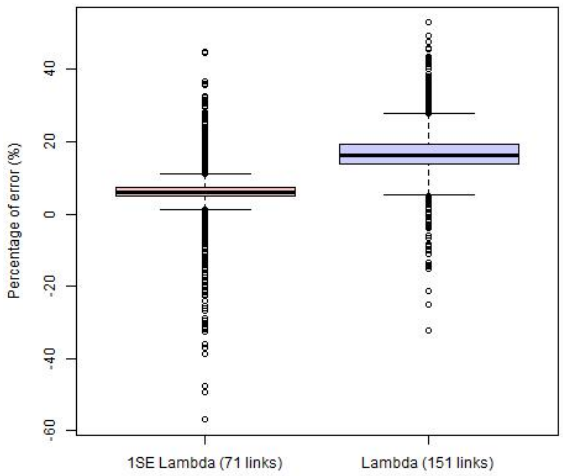
For the $NO_x$ variable, the $CO_2$ model had the lowest score followed by $CO_2 \cup NO_x$ and the lambda model. The $CO_2$ model had 77 selected links and errors around $\pm 10\%$ as the $CO_2 \cup NO_x$ model with 95 predictors and a lambda model with 174 predictors.

Following these observations, it is possible to conclude that the best models according to the $BIC$ criteria are those with at least 28% of the network links.

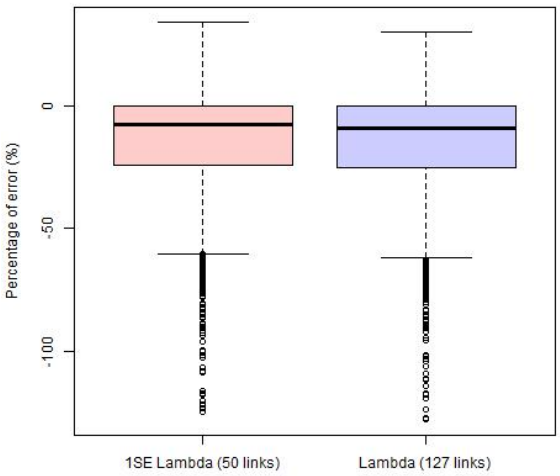
**The influence of route choice on the selection**

In contrast to the static/static dataset, all the variables from the static/dynamic dataset could be used to build a statistical model taking into account the confidence bounds. The time periods represent both daily traffic peaks: morning and afternoon. Each one represents a different scenario of route choices. Scenario 1 represents the morning peak and the scenario 2 the afternoon peak. Using scenario 1 as the training set and scenario 2 as the validation set or vice-versa, we study whether using only the data from one peak traffic period is relevant for estimating the other one. The same comparisons were made for the static/static dataset, see section 3.3.1.

Scenario 1 was represented by the $X_{n\times p}$ matrix. The $X_p$ are represented by the links in the network. The $X_{np}$ is the morning variable value of link $p^{\text{th}}$. The LASSO method was applied to this data considering the $Y$ vector which represents the value of the 15-minute period variable in the network. The number of observations is 9600. The selected model is validated in the evening (namely scenario 2). The error distribution values of both models proposed by LASSO and by variable are presented in figure 3.35. The cross-validation values for both models are presented in C.9.

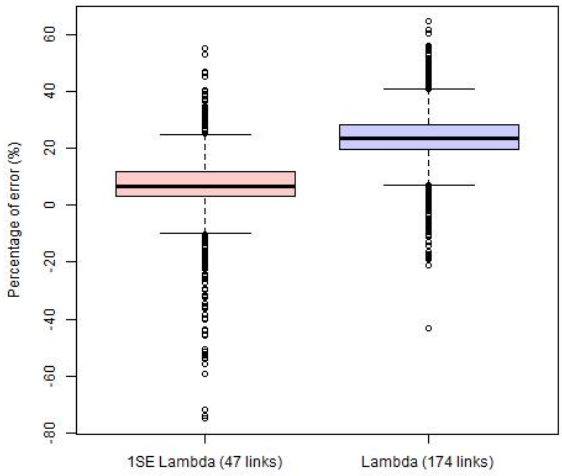



(a) Total traveled distance error distributions for lambda models.

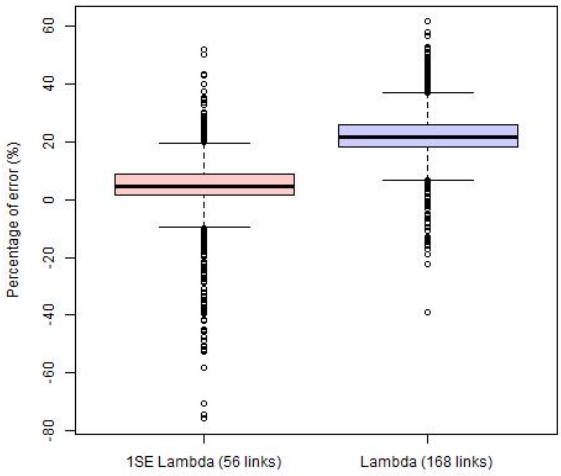



(b) Spatial mean speed error distributions for lambda models.

For the linear network variables, traveled distance and emissions, the 1SE lambda model selects between 20% to 30% of the network links. In contrast to this, the lambda model selects even more, between 65% and 75%. The last model tends to overestimate the linear variables, considering that they are twice as large compared to 1 SE lambda. On average, the lambda model overestimated the traveled distance by 17%; the $CO_2$ emissions by 24% and $NO_x$ by 22%. For these variables, the 1SE lambda model estimates better with fewer links inside the model. The error distribution of the 1SE lambda model is less scattered and, considering the outliers, the model overestimated the traveled distance by 6%, the $CO_2$ emissions by 7% and $NO_x$ by 5%. The contrary occurs for spatial mean speed. The lambda and 1SE lambda models have almost the same error distributions and both underestimate the spatial mean speed by 14% on average. Comparing both lambda models, the better




(c) Network $CO_2$ emission error distributions by lambda models.

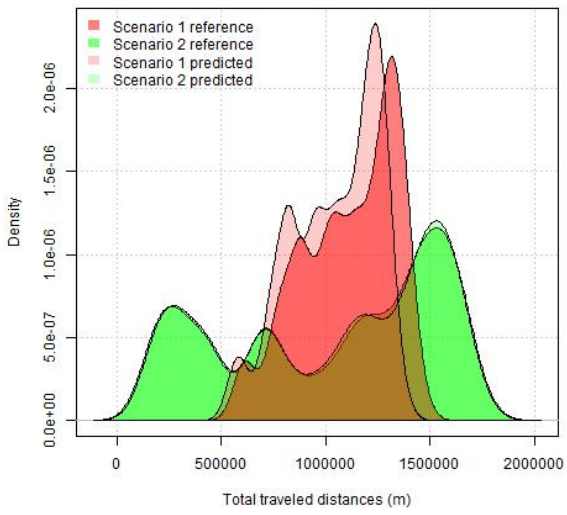



(d) Network $NO_x$ emission error distributions by lambda models.

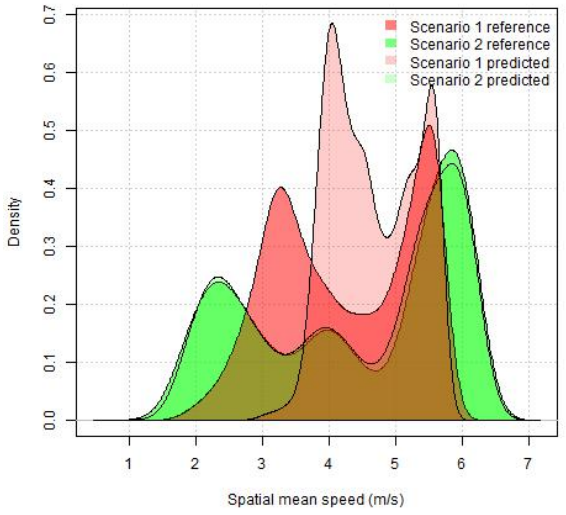
Fig. 3.35 – *Error distributions from models built using morning data and validated by the evening data.*

of the two is the 1SE lambda model for all the variables as it has fewer dispersed errors and requires more links inside the model than the lambda model.

Considering the variable values of the 1SE lambda model, the density values were analyzed to show the difference between the values estimated and the original ones. Figure 3.36 shows a comparison between four distributions. The reference values are the values that must be predicted for each scenario, knowing that the selection model was built in scenario 1 (morning peak). The predicted values are the values obtained after applying the model selection built in scenario 1 with the values of scenario 2.

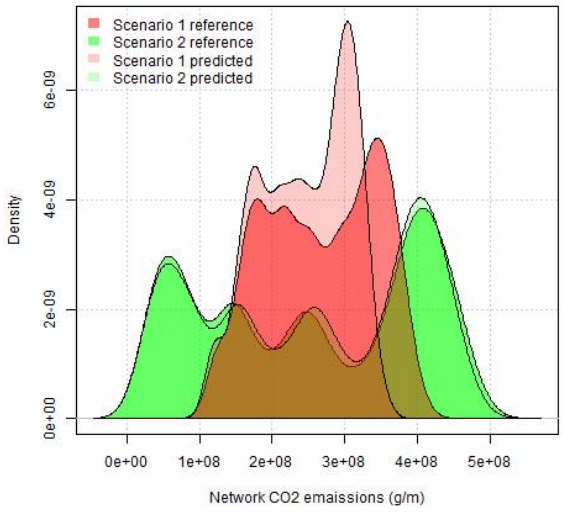



(a) Total traveled distance densities.

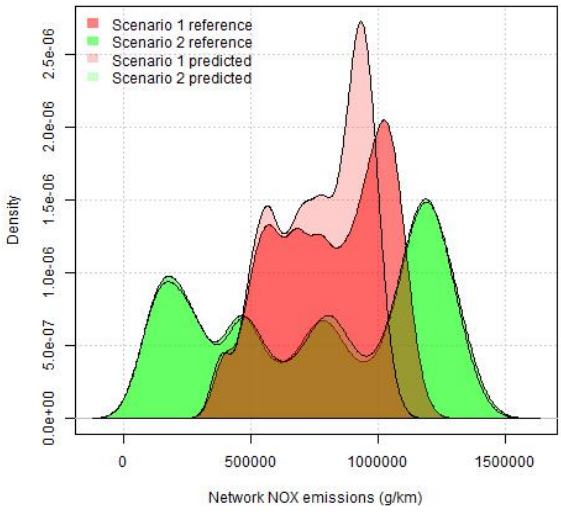



(b) Spatial mean speed densities.

In general, the scenario 1 model estimates scenario 2 better than scenario 1. This can be explained by the fact that the evening peak values are narrower than for scenario 1. These variations can




(c) Network $CO_2$ emission densities.




(d) Network $NO_x$ emission error distributions of lambda models.

Fig. 3.36 – *Density distributions for the network variables.*

be seen in section 2.1.1 in chapter 2. This model fitted to the morning peak data can estimate the evening values with an error distribution of ±20% with an average between 5% and 7% for the traveled distances and pollutant emissions. For the spatial mean speed, the values can be estimated with an error distribution between 40% and -60% with the average around -14%. The Links selected for each variable in scenario 1 are shown in 3.37.

As can be seen in figure 3.37 the selection between all the variables presents common links. The percentage of common links for each variable is shown in table 3.13.

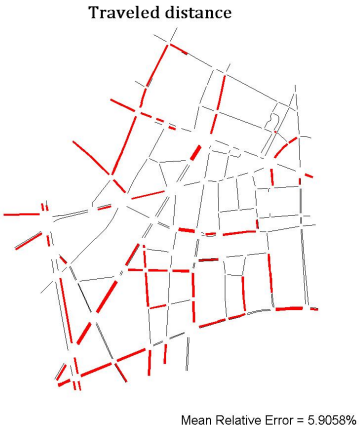
| → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| DTP | 100% | 56% | 55% | 59% |
| VIT | 39% | 100% | 36% | 39% |
| $CO_2$ | 37% | 34% | 100% | 77% |
| $NO_x$ | 46% | 44% | 91% | 100% |

Tab. 3.13 – *Common links selected in variables.*

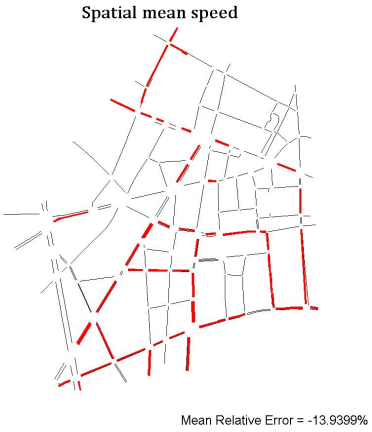
The selection between all the variables contains at least around 40% of common links. The $CO_2$ emissions selected 47 links, of which 43 are equal to those selected for $NO_x$ emissions, thus 91% of the links selected for $CO_2$ are similar to $NO_x$.

After analyzing the possibility of using the morning data to estimate the evening data, the inverse situation is proposed using the same evening data to build a model (training set) and estimate the morning data (validation set). The cross-validation of scenario 2 for all the variables is shown in C.10. The error distributions of both optimal models for each variable are shown in 3.38.

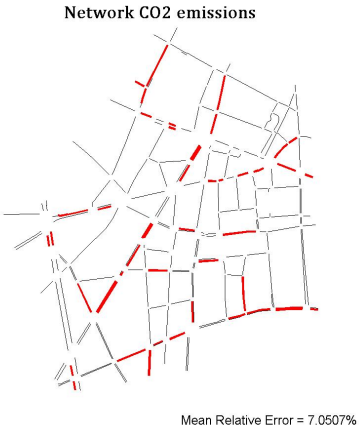
Unlike the previous case, the lambda model, which is 2.5 times larger than the SE lambda model, predicts better than the 1SE lambda model. The lambda model has a less dispersed error distribution with an average error of 1.3% for traveled distance, 7.89% for spatial mean speed, 6.7% for $CO_2$



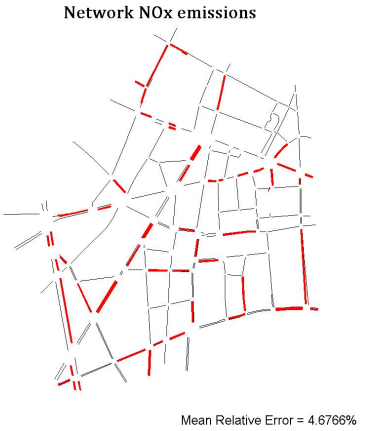

(a) Total selected traveled distance links.




(b) Selected spatial mean speed links.




(c) Selected network $CO_2$ emission links.

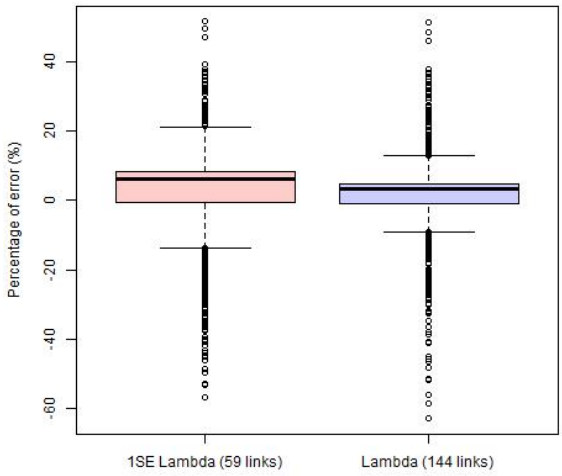



(d) Selected network $NO_x$ emission links.

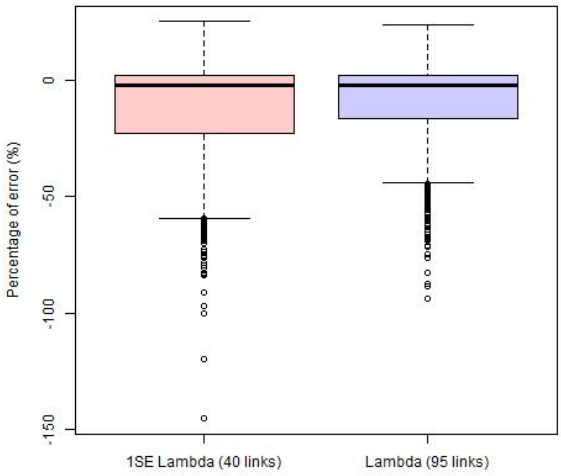
Fig. 3.37 – *Links selected by the 1 SE lambda model in scenario 1 for the network variables.*

and 3% for $NO_x$ emissions. The 1SE lambda models have a percentage error of around 11.3% on average, except for the traveled distance, which has an average percentage error of 2.5% considering all the values (including outliers). For all the models, the traveled distance and pollutant emissions are overestimated and the spatial mean speed underestimated. Thus, the same trend as in the previous study using the morning data for model fitting is observed. Even for this case, the lambda model presents better results, their model sizes are too large to use as the basis of the study. The 1 SE lambda will be chosen to analyze the quality of the prediction in greater detail. The traveled distance has 59 links inside the model, the spatial mean speed has 40 links, the $CO_2$ with 73 and the $NO_x$ emissions have 57 links. Figure 3.39 shows the distribution values of the 1SE lambda model for all the variables.

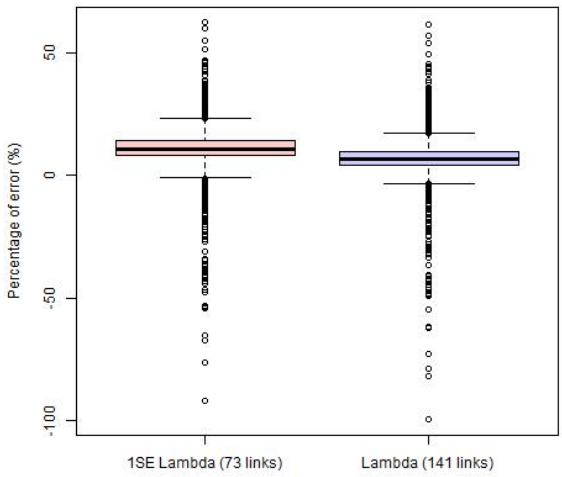
The values predicted in relation to the reference ones have the same distributions but with more visible differences between them than shown in the previous case. This explains the higher average error values using the evening data instead of the morning data to fit a model. This is understandable because the variable values are less scattered in the evening than in the morning, making adjustments to a larger set of solutions difficult. The links selected for each variable are presented in figure C.11. The percentage of common links between all the variables is slightly higher than in the previous case, highlighting the $NO_x$ model which has 79% of its links in common with $CO_2$; the same occurred for the morning case. The table with the percentage of common links is presented in C.2.



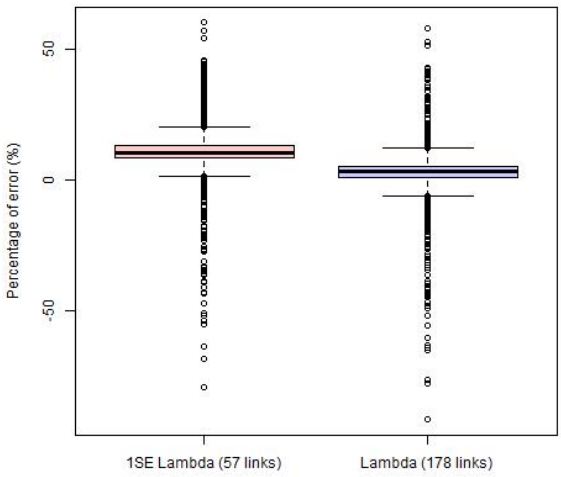

(a) Total traveled distance.




(b) Spatial mean speed.




(c) Network $CO_2$ emission.



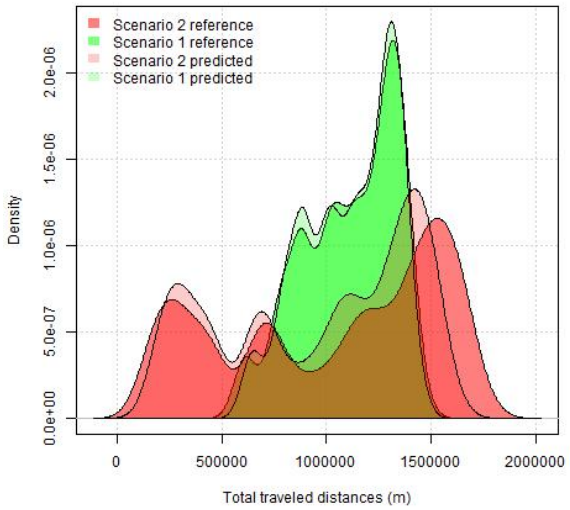

(d) Network $NO_x$ emission.

Fig. 3.38 – *Error distribution of variables.*

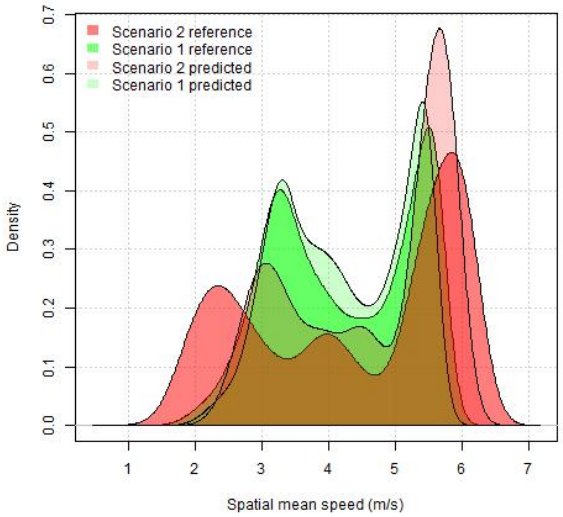
In contrast to the static/static dataset, the morning and evening peak data can be used to estimate each other. In both cases (morning and evening), the models selected slightly fewer network links with far more errors than the model applied directly to each variable at the beginning of this section. In general, comparing the morning training set leads to better results than the evening one. Analyzing each variable separately: the traveled distance and spatial mean speed are estimated better using the evening data with an average error around 2.5% and -11% respectively; scenario 1 estimated both pollutant emissions better, with an average error around 7% and 5%.

### Conclusion on the static/dynamic dataset

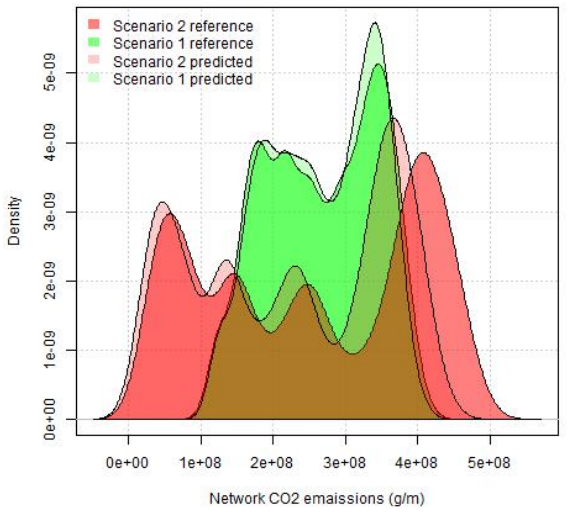
The LASSO models can reproduce the variable values with a representativeness of about 95% instead of around 70% on average for all the variables in the static/static dataset. The variance of the variable



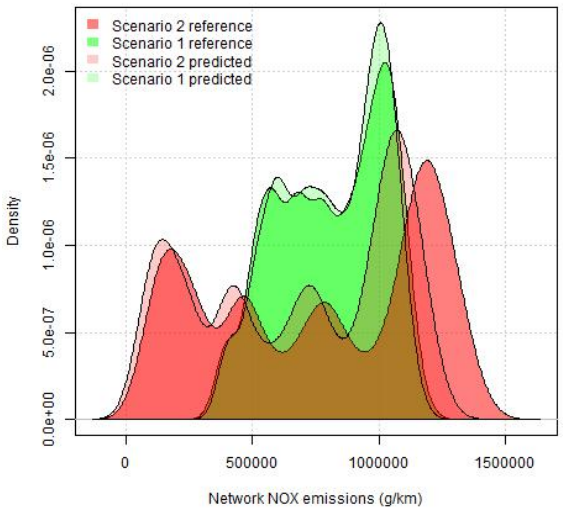

(a) Total traveled distance densities.




(b) Spatial mean speed densities.




(c) Network $CO_2$ emission desnities.




(d) Network $NO_x$ emission densities.

Fig. 3.39 – *Density values for all variable models.*

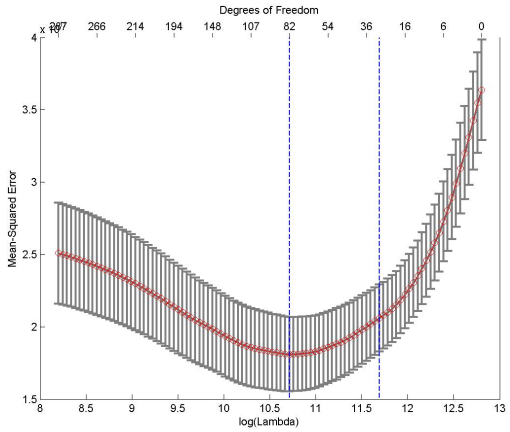
values inside this dataset is considerable due to their temporal scale which was reduced from 6 hours (static/static) to 15 minutes. This temporal scale can effectively show the free-flow and congestion periods in the network. Therefore all the models studied in this dataset presented many more outliers than the previous dataset. The outliers represent around 15% of the number of observations and they can have very high error percentage levels, which means that it is possible to estimate only 85% of the 3200 situations observed with reasonable accuracy. The models studied presented a selection rate between 28% and 78%. Only the traveled distance variable had models with an error distribution of ±5% or lower; the other variables are slightly lower than ±10%. The number of links in the model can be reduced if higher error percentage limits can be considered.

### 3.3.3 LASSO applied to dynamic/dynamic datasets

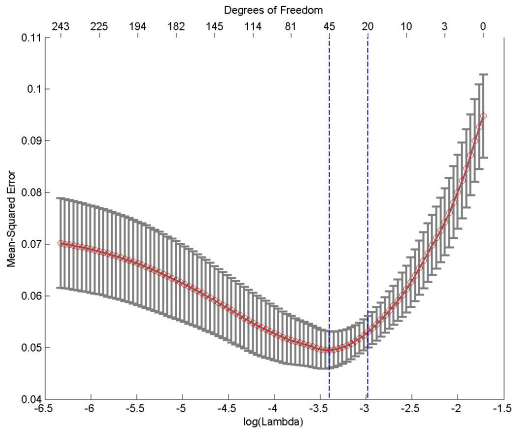
This dataset has 4 different variables: traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions. These variables are structured as a $n \times p$ matrix, where $p$ are the links represented for each period of time (230 links and 24 time periods: 5520 predictors) and $n$ are the simulation values. Each predictor has 400 observations and they were split up randomly into two parts: the first represents 2/3 of the matrix and was used as the training set to which LASSO was applied; the second part, 1/3 of the original matrix, represents the validation set against which the LASSO selection will be validated and the associated errors quantified.

#### Models proposed by LASSO

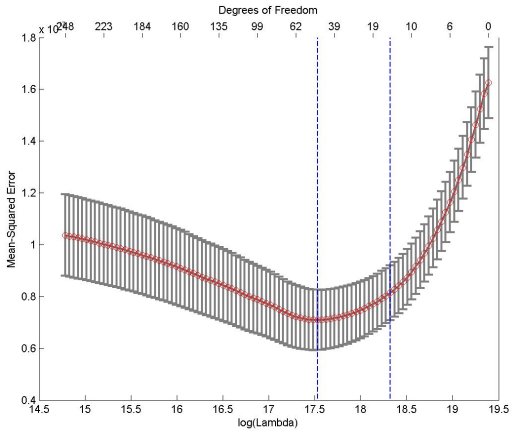
The description of this dataset organization was explained in 3.2.2. The mean-square prediction error curves for traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions are shown below.



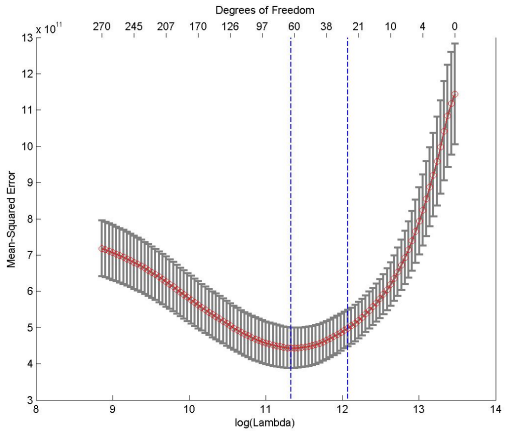

(a) Traveled distance mean-square prediction error curve.




(b) Spatial mean speed mean-square prediction error curve.




(c) $CO_2$ emissions mean-square prediction error curve.



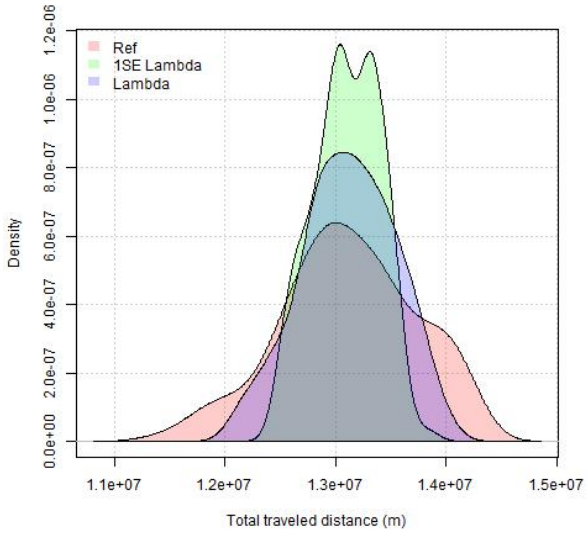

(d) $NO_x$ emissions mean-square prediction error curve.

Fig. 3.40 – *Estimated prediction error curves and their standard errors for the variables in dynamic/dynamic datasets.*

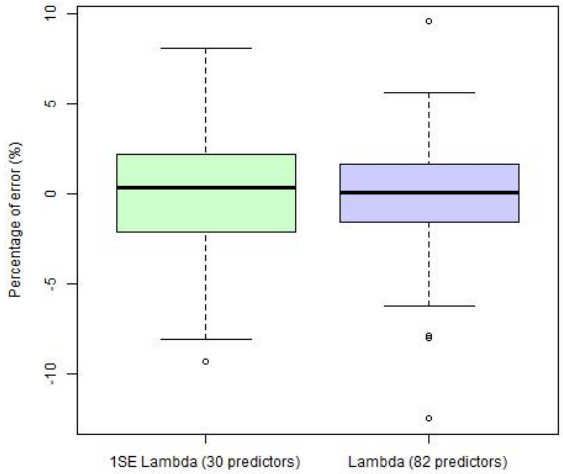
As can be seen in figure 3.40 each curve is plotted as a function of the corresponding complexity parameter ($\lambda$). The horizontal axis has been chosen so that the model's complexity increases as we move from right to left. The estimates of prediction errors and their standard errors were obtained by

tenfold cross-validation. The least complex model within one standard error is chosen from the best one and indicated by the vertical lines. The top of each plot is annotated with the sizes of the models.

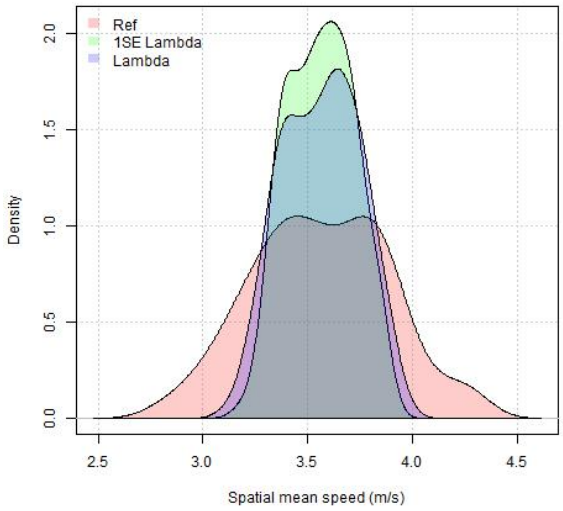
As in the other datasets, we are interested in models with fewer predictors (links and their selected period of time). To confirm the choice of the model, the resulting values of each model will be compared with the reference values and their calculated relative errors. Figure 3.41 shows the predicted values from both models for each variable and they are compared with the reference values (original values that must be predicted).



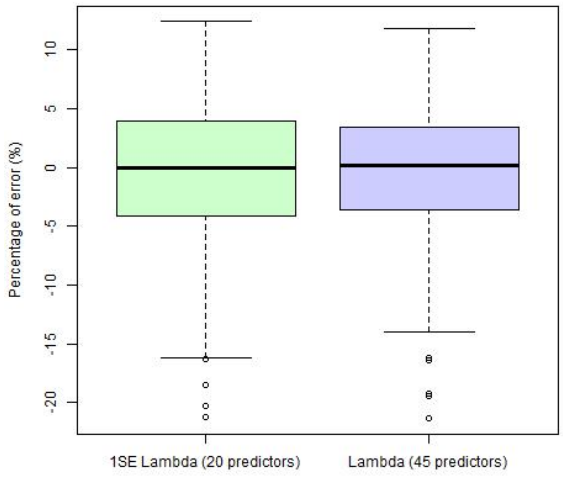

(a) Total traveled distance densities.




(b) Total traveled distance error distributions by model.

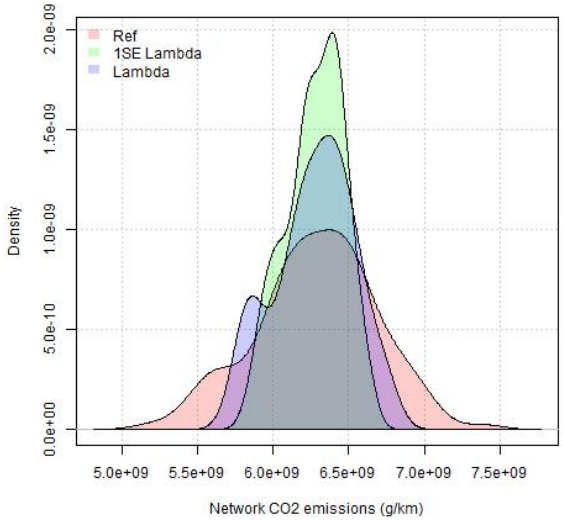



(c) Spatial mean speed densities.

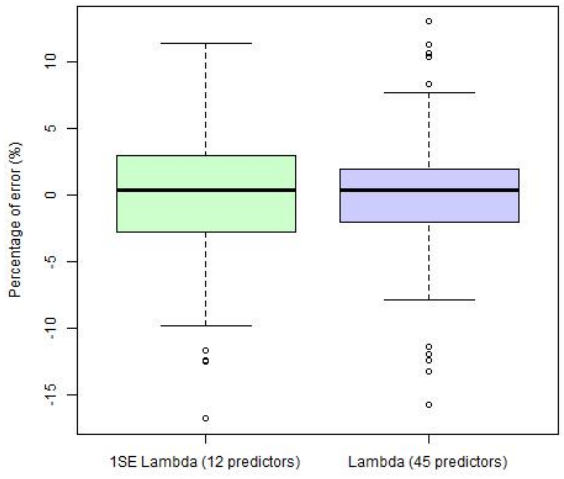



(d) Spatial mean speed error distributions by model.

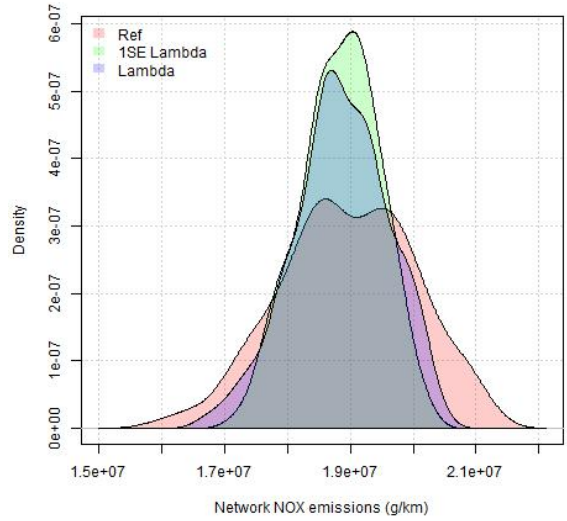
The resulting values of traveled distance using the lambda model (on the left line in 3.40) can explain 73% of the reference values versus 52% for the one standard-error model (the vertical line on the right in 3.40). For spatial mean speed the same occurs with 65% versus 56%. Regarding emissions, 70% of the data of the first optimized $CO_2$ model were explained versus 56% for the 1SE lambda model and 76% of the data of the $NO_x$ model were explained versus 54%, both compared



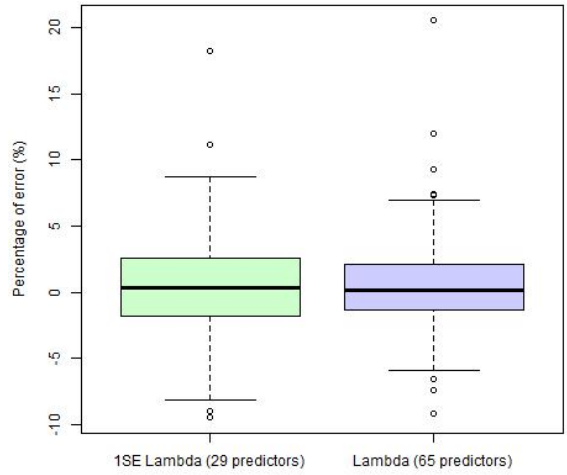

(e) Network $CO_2$ emission densities.




(f) Networ $CO_2$ emission error distributions by model.




(g) Network $NO_x$ emission densities.




(h) Networ $NO_x$ emission error distributions by model.

Fig. 3.41 – *Comparisons of models for each variable.*

with their respective reference values. The lambda model explained the data better, 15% more on average, because this model considers more predictors (links) than the 1SE lambda model and has fewer dispersed errors in comparison. For all the variables, both optimized models tended to fit models that centered the values according to the distribution of the reference values.

The associated errors for each model using the reference values were also calculated and the results are shown in 3.41. Both models have similar error distributions considering all the variables. The error distribution of traveled distance is lower than ±10%; the spatial mean speed presents an error between slightly more than -15% and 10%; the percentage error of the $CO_2$ emissions is around ±10%; and, finally, $NO_x$ emissions present an error percentage lower than ±10%.

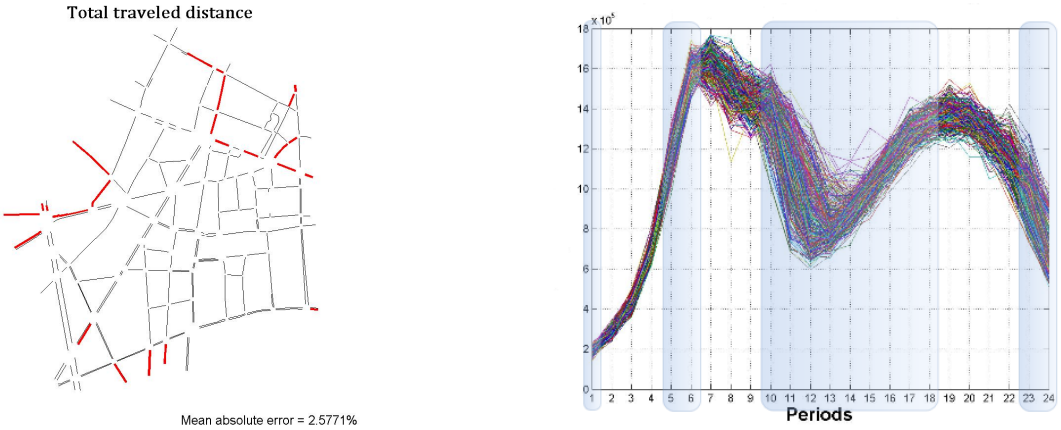
If the size of the model of each variable is considered, the model defined by the one standard-error rule from lambda is the best because it considers far fewer predictors for the same average error and

data representativeness.

### The 1SE model

As explained before, the 1SE lambda model was the model selected for all the variables. The results are presented only for this model. In figure 3.40 are shown the lambda cross-validated values for each variable.

Regarding traveled distance, the 30 predictors selected represent 25 links of the network. 14 predictors correspond to morning traffic versus 16 in the evening (the morning traffic is represented by periods of time from 1 to 12 and the evening from 13 to 24). It is interesting to observe that the selected time periods mostly correspond to free-flow periods. These datasets provide as results the daily values of the network as in the static/static datasets. The model built for traveled distance can explain 52% of the data (with a 95% confidence interval) and gives results with a mean absolute error equal to 3.6%. If we compare the traveled distance selection between the static/static and the dynamic/dynamic datasets, we have 6 common links, while 6/7 links selected for traveled distance in static/static datasets are included in the model built by LASSO in the dynamic/dynamic dataset.



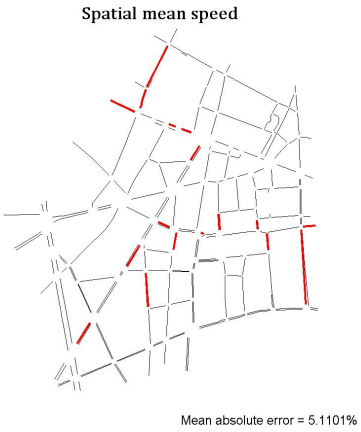

(a) Selected traveled distance links.

(b) Selected traveled distance time periods.

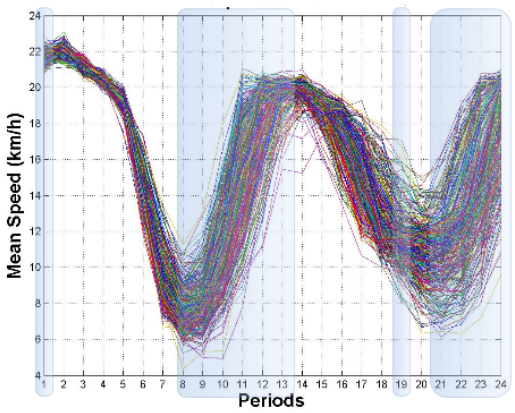
Fig. 3.42 – *Traveled distance selection.*

The spatial mean speed had 20 time periods selected, which represents 18 links. Of the 20 predictors, 12 are from the morning peak and 8 are from the evening. The model can explain 56% of the data with an average absolute error of 5.11%. Most of the periods selected represent a congested state in the network, in complete opposition to the traveled distance. This could be explained by the opposing correlation between both variables. This dataset gives us the average daily speed in the network as a result. If we compare it with the selection made in the static dataset (which has 19 links in the 1SE lambda model), they have only 6 in common. The selection made by LASSO is shown in figure 3.43.

Figure 3.44 represents the selection made in $CO_2$ emissions: 12 predictors are selected. They represent 11 links in the network, 5 of the selected time periods are in the morning assignment, 7 in the evening peak, 8 predictors out of 12 are the same (link and time period) between $CO_2$ and traveled distance. None of the predictors selected for spatial mean speed were selected for $CO_2$. The periods selected represent mostly free-flow state in the network. The model built by LASSO can explain 56% of the data with an average absolute error of about 3.6% as shown in figure 3.46.

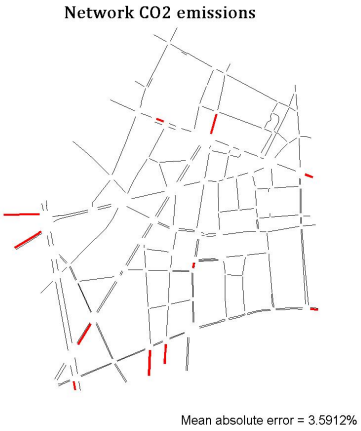



(a) Selected spatial mean speed links.

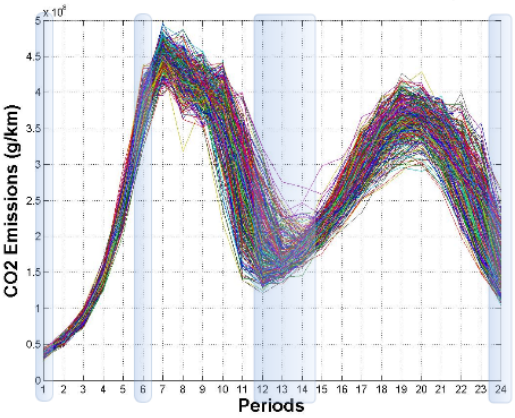



(b) Selected time periods for spatial mean speed.

Fig. 3.43 – *Spatial mean speed selection.*




(a) Selected $CO_2$ emission links.

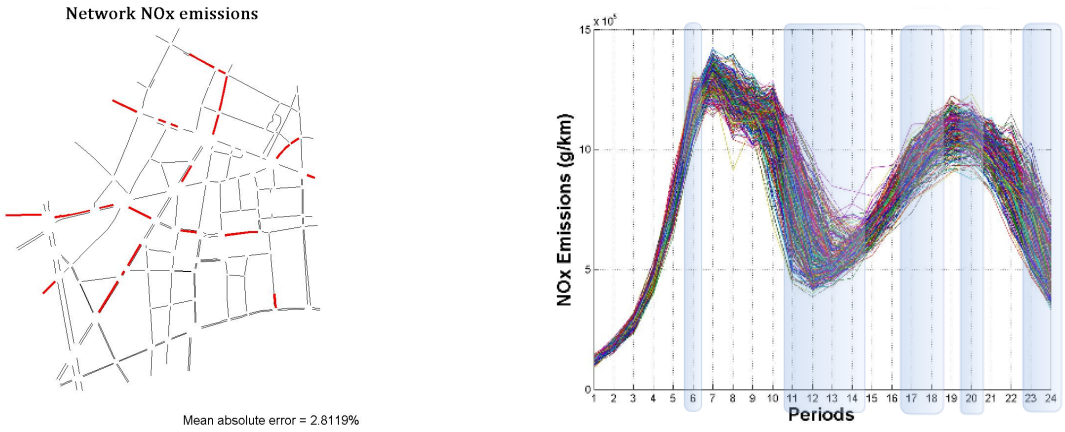



(b) Selected time periods in $CO_2$ emissions.

Fig. 3.44 – $CO_2$ *emission selection.*

Figure 3.45 presents the selection made for the $NO_x$ emissions. This variable had 29 selected predictors and they are represented by 23 links with 6 predictors for the morning traffic and 23 predictors for the evening peak. This variable selected more evening time periods than the other variables in this dataset. The model built by LASSO using the selected links and time periods can explain 64% of data with an average absolute error of 2.8%.

When analyzing the results of all the variables, free-flow time periods are selected for most of them, considering the linear variables as traveled distance and pollutant emissions, which means traffic states in which the network is loaded with vehicles but with normal traffic flows. Observations to the contrary can be made for spatial mean speed, for which most of the congested time periods were selected. The percentage of selected time periods is generally higher in the evening than in the morning except for spatial mean speed which selected most of the periods in the morning traffic. Traveled distance selected 47% of time periods in the morning and 53% in the evening; spatial mean speed selected 60% in the morning and 40% in the evening; $CO_2$ selected 42% and 58% in morning and evening respectively; and finally $NO_x$ emissions selected 21% of the predictors in the morning peak and 79% in the evening. The emission models selected completely different links and time periods in comparison to the selection obtained with spatial mean speed.




(a) Selected $NO_x$ emission links.

(b) Selected time periods in $NO_x$ emissions.

Fig. 3.45 – *The selected time periods and links for $NO_x$ emissions.*

The daily values of each variable were calculated using their respective validation sets and their relative errors were calculated. The error distributions are almost the same as for the other datasets, 50% of the data have a percentage error lower than 7%. The linear variables present an error distribution lower than ±10% and that of the spatial mean speed is between -20% and 15%. In figure 3.46, the error distributions of all the variables are shown.

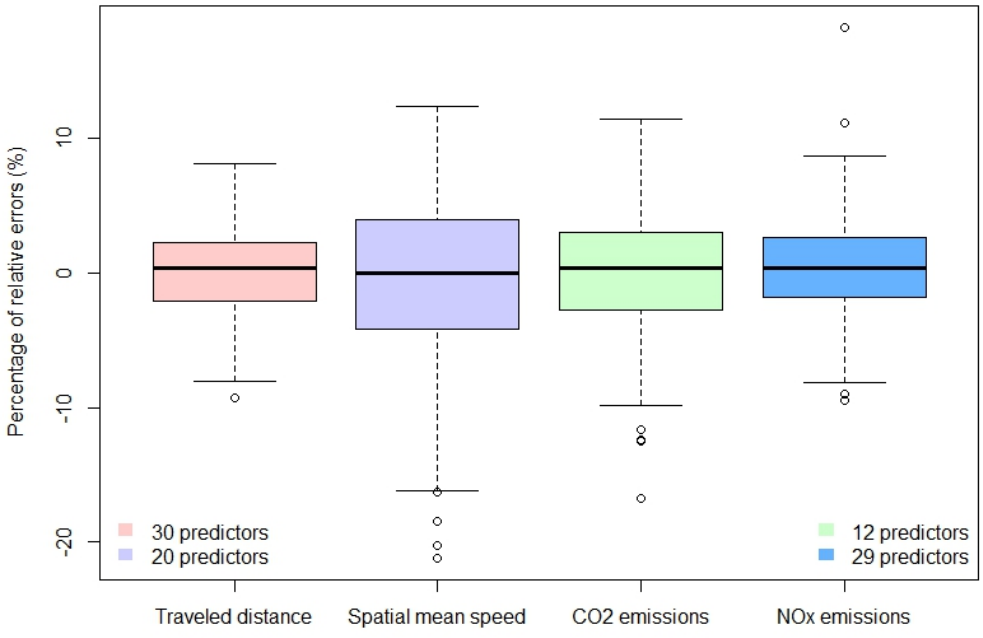



Fig. 3.46 – *Error percentage between the predicted variable values using the 1SE lambda model and the original values of the variables.*

Using slightly more links than the static/static dataset and a lot fewer than the static/dynamic dataset, it is possible to estimate the network values with the same range of error as both the previous datasets using fewer data.

### Model size and error distribution

As studied previously for the other datasets, the number of predictors inside the model versus their error distributions will be studied to determine the best compromise. Only the main figures will be presented in this section; the others can be found in appendix D.1.

Figure D.1 presents the possible models that have fewer selected predictors than in 1 SE Lambda model (*). Axis x shows the number of predictors in each model and not the number of links selected in the network. The range of selected links in these models is given as follows: the model with 3 predictors has 3 links, the 1 SE lambda (*) model has 30 predictors that correspond to 25 links and finally the lambda (**) model with 82 predictors is represented by 64 links.

The error distributions fall slightly as the number of predictors increases, until reaching the 1 SE lambda model. The error distribution of the lambda model does not fall within the range of ±5%. Consequently, the model with more predictors will be investigated in figure 3.47.

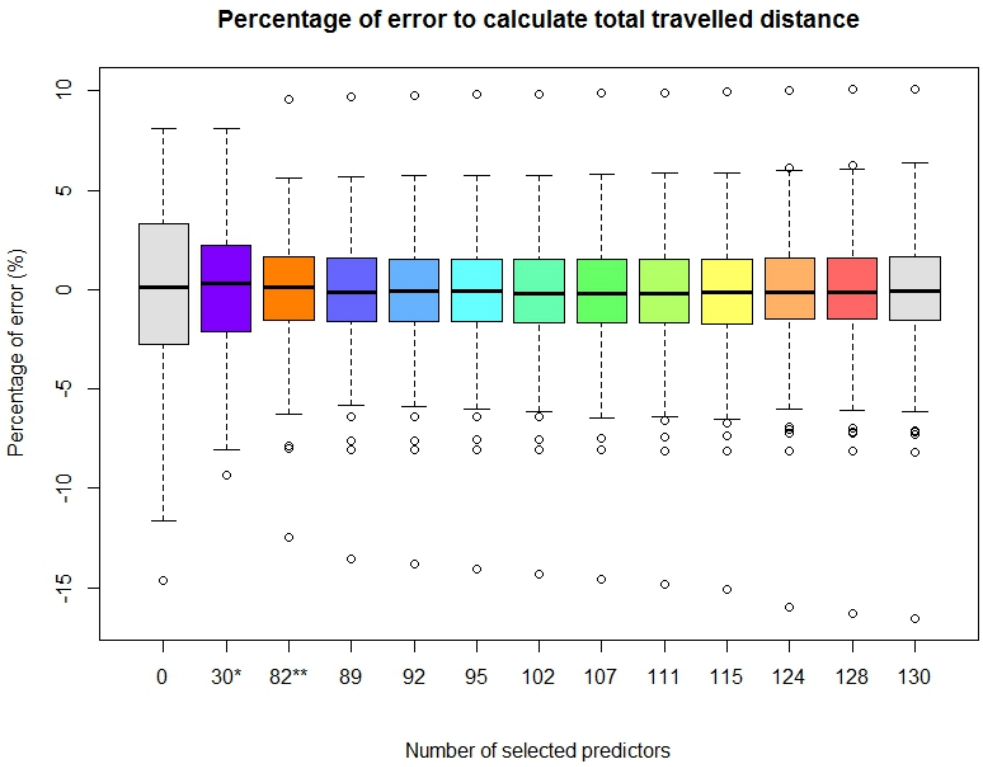



Fig. 3.47 – *Error distributions versus number of predictors to estimate network traveled distance.*

The error distributions between all the models do not change markedly, even when increasing the number of selected predictors. The model with 130 predictors corresponds to 86 links of the network and finally has almost the same error as the lambda model with 82 predictors (64 links).

The errors are sparser for spatial mean speed than for traveled distance. Figure D.2 shows the error distributions using the LASSO method to reduce the number of selected predictors as a function of the 1 SE Lambda model. As observed for traveled distance, the error distributions are not reduced significantly when the number of predictors is increased. In this figure the model with 2 predictors corresponds to 2 links of the network, and each link corresponds to a time period. The 1 SE lambda model selected 20 predictors which correspond to 18 links and, finally, the lambda model selected 45 predictors with 34 links.

Considering the error distributions of the lambda model are higher than ±10%, the models that

selected more predictors than the lambda model are presented in figure 3.48 to observe whether this error distribution is reduced.

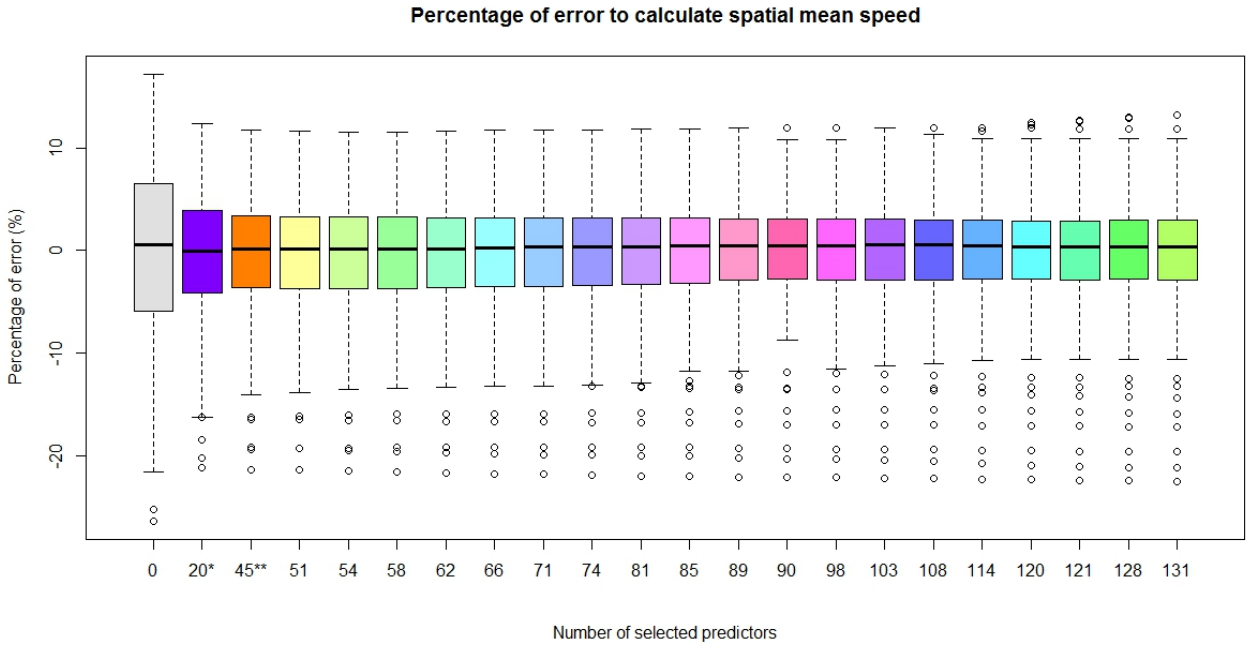



Fig. 3.48 – *Error distributions versus number of predictors to estimate spatial mean speed.*

Although 3 times the size of the lambda model, the error distributions continue around ±10% and present more outliers than models with fewer predictors. The model with 131 predictors corresponds to 97 links, so more than 40% of the network.

The LASSO model for $CO_2$ contains 12 predictors (11 links) for the 1SE model and 45 predictors (37 links) for the lambda model. Figure D.3 shows the LASSO models used to estimate $CO_2$ emissions with fewer predictors than the 1 SE lambda model. The error distributions of most of the models are between ±15% and ±10%. The percentage error of the lambda model is lower than ±10%. The models with more predictors than the lambda model are shown in figure 3.49.

The error distributions do not change that much considering the higher number of predictors in the models. Only the outliers tend to have a higher number of errors than the model just before them. The last model with 133 predictors contains 98 links. The model with 97 predictors is less dispersed than the others.

The last variable to be analyzed is $NO_x$ emissions. Like the $CO_2$ emissions, the number of errors in the models with fewer predictors than the 1 SE lambda model falls smoothly as the number of predictors within each model increases. The 1SE lambda model (*) has 29 predictors with 23 links while the lambda model has 65 predictors that represent 48 links. The percentage error is about ±10% excluding the outliers from the model with 15 predictors up to the lambda model. All these conclusions can be observed in figure 3.50.

The error distributions of the models with more predictors than the lambda model are similar, even when considering their outliers. The number of errors in the model with 139 predictors (94 links) is similar to that with 65 predictors (48 links). This trend was observed for all the variables. The models ranging from the lambda model to the last model have similar errors despite the fact that they use more link information than the previous model. The lambda model for all the variables can be considered as the best choice when the number of predictors and error distributions are compared.

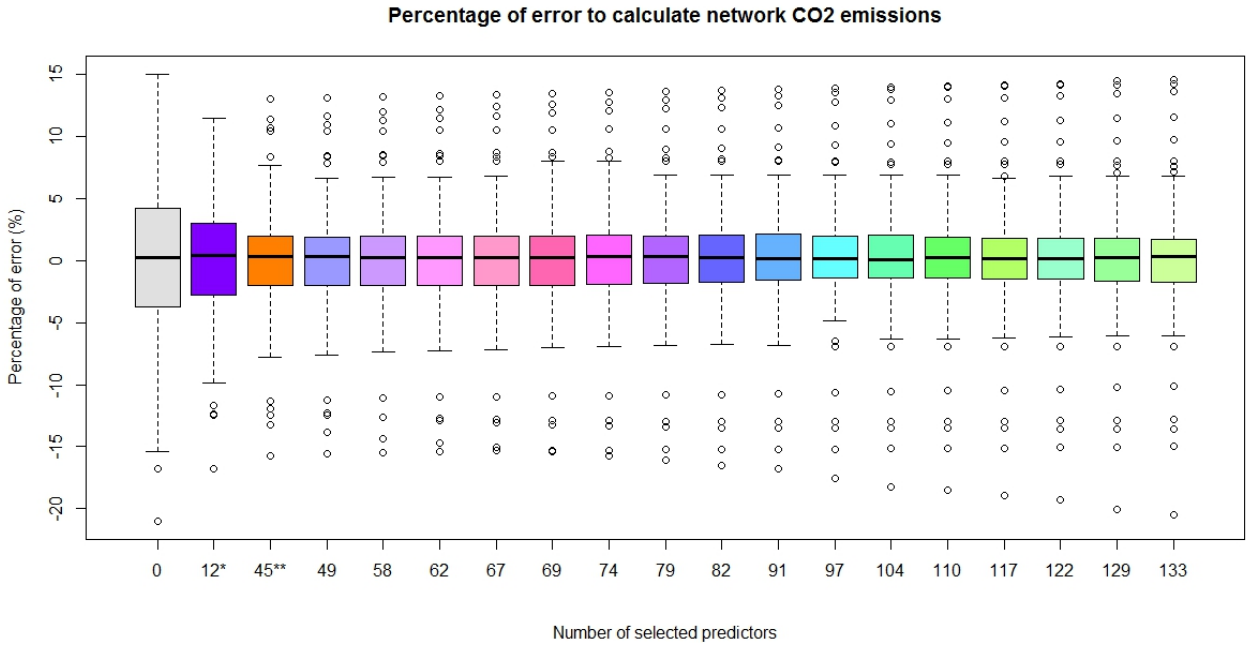



Fig. 3.49 – *Error distributions versus number of predictors to estimate network $CO_2$ emissions.*

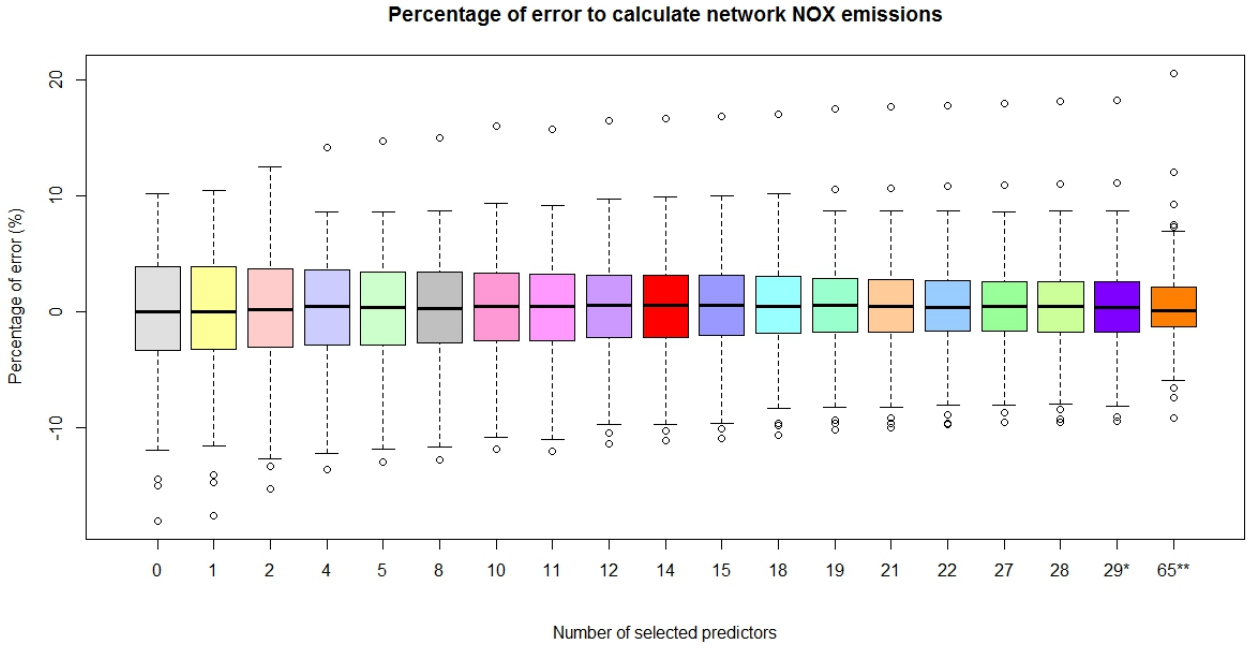



Fig. 3.50 – *Error distributions versus number of predictors to estimate network $NO_x$ emissions.*

Until now, the study has been conducted to evaluate variables according to their number of selected links in the network and their error distributions in their models. The aim of the dynamic/dynamic dataset is to select the most pertinent time periods for estimating the network variables. Free-flow zones and congested periods are defined based on figures 3.42(b) and 3.43(b) which show the total traveled distance of the entire network and the spatial mean speed for each time period and simulation. Table 3.14 shows the percentage of selected periods in the models presented in the study previously such as: minimum number of selected predictors, the 1 SE lambda model, the lambda model, and the model with the maximum number of selected predictors.

Several observations were taken into account before analyzing each variable separately: time period

number 12 was selected in all the cases presented and always with a representative percentage of selection; after that, the time periods around 12 such as 11 and 13, and also 23 and 24, were those selected most when considering all the models. The number of predictors in the models increased followed the trend of selecting uniformly distributed time periods. It is important to say that time period number 12 is the period of time that has the highest correlation value for all the variables, as shown in chapter 2.

Starting with traveled distance, all of the four cases analyzed selected more free-flow periods than congestion periods. For the spatial mean speed, the first 3 cases had more congested periods selected than those in free flow. The model with 131 predictors had almost the same percentage of selection between both traffic states. Considering both pollutant emissions and excluding their first model with 2 predictors for $CO_2$ and only one for $NO_x$, all the time periods selected for the models mostly fell in free-flow traffic state.

Considering all these conclusions, the correlations presented in chapter 2 explain the percentage of time periods selected. Regarding the strong correlation between the variables and considering that it is positive, the LASSO method tends to select the time periods in free flow state, as shown for traveled distance, $CO_2$ emissions and $NO_x$ emissions. When the strong correlations are inverted, as with spatial mean speed, the method tends to select the time periods that are congested.

Tab. 3.14 – *Percentage of selected time periods for each model and variable.*

| VARIABLES → | | DTP | | | | VIT | | | | $CO_2$ | | | | $NO_x$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PERIODS ↓ | Traffic state | 3 selected predictors (3 links) | Lambda 1SE: 30 predictors (25 links) | Lambda min: 82 predictors (64 links) | 130 selected predictors (86 links) | 2 selected predictors (2 links) | Lambda 1SE: 20 predictors (18 links) | Lambda min: 45 predictors (34 links) | 131 selected predictors (97 links) | 2 selected predictors (2 links) | Lambda 1SE: 12 predictors (11 links) | Lambda min: 45 predictors (37 links) | 133 selected predictors (98 links) | 1 selected predictors (1 links) | Lambda 1SE: 29 predictors (23 links) | Lambda min: 65 predictors (48 links) | 139 selected predictors (94 links) |
| *1* | *Free-flow* | 66,67% | 20,00% | 10,98% | 8,46% | - | 5,00% | 2,22% | 2,29% | - | 16,67% | 8,89% | 6,02% | - | - | 7,69% | 3,60% |
| *2* | *Free-flow* | - | - | - | 0,77% | - | - | - | 2,29% | - | - | - | 3,76% | - | - | - | 0,72% |
| *3* | *Free-flow* | - | - | 1,22% | 6,15% | - | - | - | 0,76% | - | - | - | 2,26% | - | - | 3,08% | 4,32% |
| *4* | *Free-flow* | - | - | 9,76% | 6,92% | - | - | - | 3,05% | - | - | 2,22% | 3,01% | - | - | - | 2,16% |
| *5* | *Free-flow* | - | 3,33% | 8,54% | 6,15% | - | - | - | 3,82% | - | - | - | 3,01% | - | - | 4,62% | 0,72% |
| *6* | *Free-flow* | - | 6,67% | 8,54% | 6,92% | - | - | 2,22% | 3,05% | - | 8,33% | 6,67% | 3,76% | - | 3,45% | 3,08% | 5,04% |
| *7* | *Congested* | - | - | 3,66% | 4,62% | - | - | 2,22% | 3,05% | - | - | - | 6,02% | - | - | 1,54% | 6,47% |
| *8* | *Congested* | - | - | - | 1,54% | - | 5,00% | 6,67% | 6,11% | - | - | - | 1,50% | - | - | 1,54% | 2,88% |
| *9* | *Congested* | - | - | 1,22% | 3,85% | - | 10,00% | 4,44% | 3,05% | - | - | - | 5,26% | - | - | - | 3,60% |
| *10* | *Congested* | - | - | 1,22% | 4,62% | - | 5,00% | 11,11% | 8,40% | - | - | - | 2,26% | - | - | - | 0,72% |
| *11* | *Congested* | - | 6,67% | 7,32% | 6,92% | - | 10,00% | 8,89% | 6,87% | - | - | 8,89% | 4,51% | - | 6,90% | 4,62% | 3,60% |
| *12* | *Congested* | 33,33% | 10,00% | 9,76% | 10,00% | 50,00% | 25,00% | 15,56% | 6,87% | 100,00% | 16,67% | 8,89% | 7,52% | 100% | 10,34% | 4,62% | 3,60% |
| *13* | *Congested* | - | 16,67% | 1,22% | 3,85% | 50,00% | 5,00% | 6,67% | 6,11% | - | 25,00% | 8,89% | 3,01% | - | 10,34% | 6,15% | 1,44% |
| *14* | *Free-flow* | - | 3,33% | 3,66% | 3,08% | - | - | - | 3,05% | - | 8,33% | 6,67% | 5,26% | - | 6,90% | 4,62% | 4,32% |
| *15* | *Free-flow* | - | 6,67% | 3,66% | 2,31% | - | - | - | 1,53% | - | 8,33% | 6,67% | 4,51% | - | - | 4,62% | 7,19% |
| *16* | *Free-flow* | - | 6,67% | 3,66% | 2,31% | - | - | - | 3,05% | - | - | 6,67% | 4,51% | - | 6,90% | 6,15% | 5,76% |
| *17* | *Free-flow* | - | 3,33% | 3,66% | 3,85% | - | - | 2,22% | 1,53% | - | - | 11,11% | 9,02% | - | 10,34% | 12,31% | 8,63% |
| *18* | *Free-flow* | - | 3,33% | 4,88% | 4,62% | - | - | 2,22% | 3,05% | - | - | 2,22% | 2,26% | - | 3,45% | 4,62% | 5,76% |
| *19* | *Free-flow* | - | - | 3,66% | 1,54% | - | 10,00% | 4,44% | 6,11% | - | - | 4,44% | 5,26% | - | - | 3,08% | 2,88% |
| *20* | *Congested* | - | - | 1,22% | 1,54% | - | - | - | 1,53% | - | - | - | 2,26% | - | 3,45% | 3,08% | 7,91% |
| *21* | *Congested* | - | - | 1,22% | 3,08% | - | 5,00% | 6,67% | 4,58% | - | - | - | 3,76% | - | - | - | 0,72% |
| *22* | *Free-flow* | - | - | - | - | - | 5,00% | 11,11% | 6,11% | - | - | - | 0,75% | - | - | 1,54% | 3,60% |
| *23* | *Free-flow* | - | 3,33% | 7,32% | 5,38% | - | 10,00% | 8,89% | 6,87% | - | - | 8,89% | 6,02% | - | 17,24% | 4,62% | 4,32% |
| *24* | *Free-flow* | - | 10,00% | 3,66% | 1,54% | - | 5,00% | 4,44% | 6,87% | - | 16,67% | 8,89% | 4,51% | - | 20,69% | 18,46% | 10,07% |
| Free-flow rate selection | | 66,67% | 66,67% | 73,17% | 60,00% | 0,00% | 35,00% | 37,78% | 53,44% | 0,00% | 58,33% | 73,33% | 63,91% | 0,00% | 68,97% | 78,46% | 69,06% |
| Congested rate selection | | 33,33% | 33,33% | 26,83% | 40,00% | 100,00% | 65,00% | 62,22% | 46,56% | 100,00% | 41,67% | 26,67% | 36,09% | 100,00% | 31,03% | 21,54% | 30,94% |

### The robustness of the models

As with the other datasets, a cross-analysis was conducted to observe if one of the 4 models, one for each variable, could be used to estimate the other variables values. Table 3.15 shows the average percentage of absolute errors in the validation sets of the 1SE lambda model on the variables in the lines and applied to the variables arranged in the columns.

| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 25 | 10,87% | 2,40% | 2,58% | 5,14% | 5,88% | 2,44% | 3,02% | 2,15% | 2,47% |
| VIT | 18 | 7,83% | 2,42% | 2,72% | 4,38% | 5,11% | 2,96% | 3,41% | 2,64% | 3,01% |
| $CO_2$ | 11 | 4,78% | 2,08% | 2,24% | 5,79% | 6,26% | 3,30% | 3,59% | 2,46% | 2,49% |
| $NO_x$ | 23 | 10,00% | 2,42% | 2,72% | 3,35% | 4,20% | 2,96% | 3,41% | 2,55% | 2,81% |
| Number of observations | | | *267* | *133* | *267* | *133* | *267* | *133* | *267* | *133* |

Tab. 3.15 – *The average absolute error of the 1 SE lambda model selection on one variable applied on another.*

As noted in the table, the same considerations for the static/static dataset are applied here. It is possible to use the selected links of one variable to determine the other ones with the same accuracy as LASSO did. Unlike the spatial mean speed, the other three variables have predictors (link and time period) in common and are shown in table 3.16. The most important conclusion is that the model defined for $CO_2$ emissions has 75% of same selected links and time periods when compared with traveled distance, which indicates strong inter-dependency.

| MODELS ↓ | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| DTP | 100% | 0% | 30,0% | 26,7% |
| VIT | 0% | 0% | 0% | 0% |
| $CO_2$ | 75,0% | 0% | 100% | 33,3% |
| $NO_x$ | 27,6% | 0% | 13,8% | 100% |

Tab. 3.16 – *The common selected links and time periods between variables.*

### Study of the merger and intersection between variables of the network

A second study was conducted to observe if a linear regression model including a set of selected predictors could be used to estimate all the variable values. We considered the merged traffic variables, the merged emissions, and the intersections on the latter. Table 3.17 presents the average percentage of absolute error between the linear regression model and the respective reference values in the validation set.

The size of the linear regression model corresponds to the number of selected predictors of two variables of the same nature, traffic and emissions. Each predictor corresponds to a singular link and time period. All the linear regression models applied to all the variables have low average percentage errors in the validation set. The size of the model with the merged traffic variables was equal to 50 predictors, corresponding to 43 links of the network considering a mixture between free-flow periods

taken from traveled distance and congested periods taken from spatial mean speed.

| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTPuVIT | 43 | 18,70% | 1,67% | 2,56% | 3,27% | 4,44% | 2,34% | 2,86% | 1,99% | 2,43% |
| DTP∩VIT | - | - | - | - | - | - | - | - | - | - |
| $CO_2$u$NO_x$ | 30 | 13,04% | 1,84% | 2,10% | 5,00% | 5,39% | 2,26% | 2,72% | 1,87% | 2,62% |
| $CO_2$∩$NO_x$ | 3 | 1,30% | 2,26% | 2,51% | 6,97% | - | 2,98% | 3,31% | 2,68% | 2,76% |
| Number of observations | | | *267* | *133* | *267* | *133* | *267* | *133* | *267* | *133* |

Tab. 3.17 – *The average percentage absolute error of the linear regression model fitted to the merge and intersection between variables of the same type.*

The merge between the two pollutant emissions is composed of 37 selected predictors over the 5520 from the original matrix. Its model corresponds to 30 selected links considering most periods in free-flow state. The last linear regression model, the intersection between $CO_2$ and $NO_x$, has only 4 predictors that represent 3 links of the network with all of them in free-flow state. The latter model can accurately assess, with 15 minute traffic data on only 3 links identified by LASSO, the daily values of the network for traveled distance and pollutant emissions. The model can explain about 60% of the data, with a confidence interval of 95%. The linear regression on the intersection between selected links is not statistically representative for the spatial mean speed.

### The best model for each variable

To evaluate which model can best estimate a single variable, the error distribution from each one is shown with the respective model sizes. Figure 3.51 shows the error distributions for all the models used to estimate the network traveled distance.

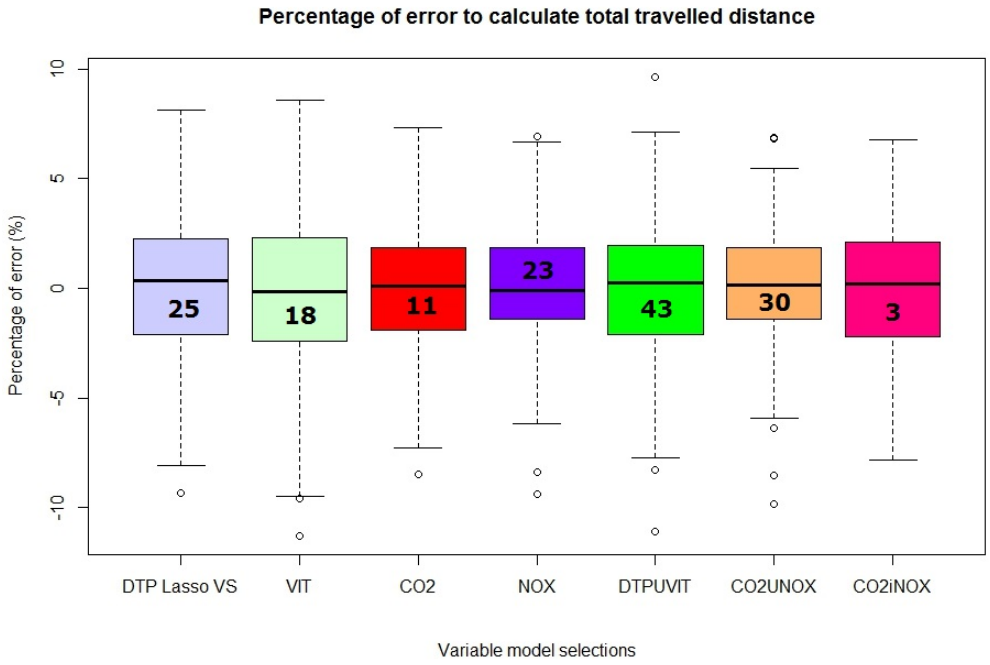



Fig. 3.51 – *Error distributions for models used to estimate the network traveled distance and the number of selected links in each model.*

The least dispersed model is the $CO_2 \cup NO_x$ containing 30 links with 37 predictors and a percentage error slightly higher than ±5%. All the models presented errors below ±10%. If the error considered for choosing a model must be lower than ±10%, this leads to a model size of $CO_2 \cap NO_x$ equal to 4 and only 3 selected links. The $CO_2$ model presents the same error distribution as the $CO_2 \cap NO_x$ with 12 predictors which represent 11 links in the network.

Figure 3.52 represents the models estimating spatial mean speed. All the models have more dispersed error distributions than for the traveled distance variable. The model that best estimates spatial mean speed is $DTP \cup VIT$ because it is the only one whose errors are around ±10% compared to the others, though with a model size equal to 43 links (19% of the network).

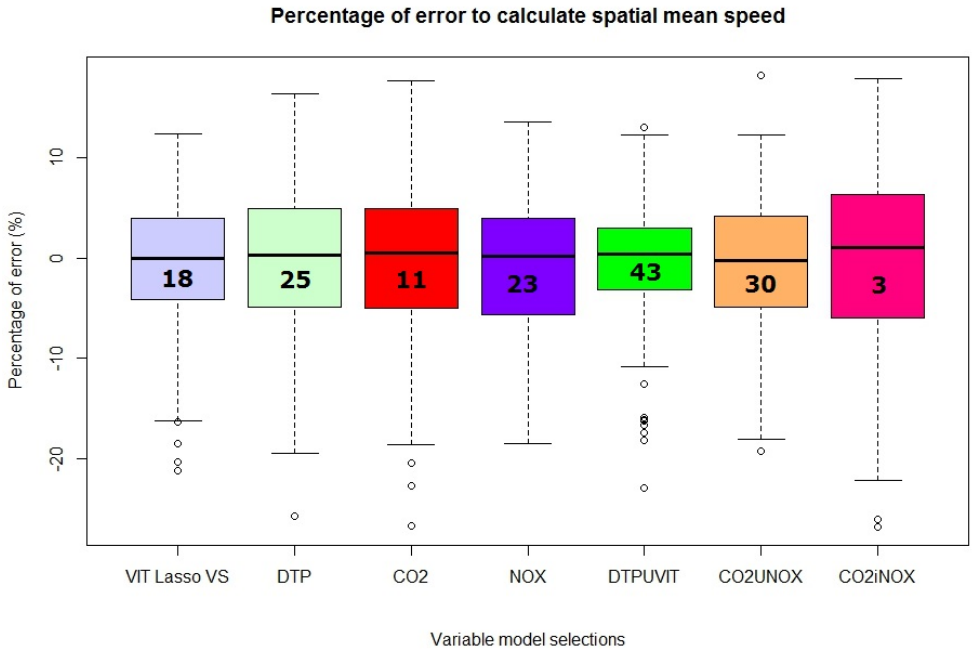



Fig. 3.52 – *Error distributions of models used to estimate the spatial mean speed and the number of selected links inside each model.*

For the $CO_2$ variable, the $NO_x$ model had the lowest amount of scattered errors with 29 predictors corresponding to 23 links in the model. Most of the models had errors distributed between -10% and 10% and certain outliers with an percentage error over 10%. These conclusions are shown in figure 3.53.

Considering the $NO_x$ variable two models have almost the same error distributions: the $CO_2$ and $DTP \cup VIT$. Their error distributions are slightly higher than ±5%, but they have different model sizes. The first has 12 predictors which correspond to 11 links of the network; the second has 50 predictors which correspond to 43 links. Excluding the outliers, all the models present errors between -10% and 10%.

All the models established by LASSO in this study can estimate other variables. The difference between them is the number of selected links and time periods. The time periods have been described and studied in this section, showing a trend by which most of the selected time periods are taken from free-flow state for the variables (traveled distance and pollutant emissions) and from free-flow for the spatial mean speed.

The $BIC$ (Bayesian information criteria) are provided in table 3.18 for all the models considered and for each variable in the dynamic/dynamic datasets.

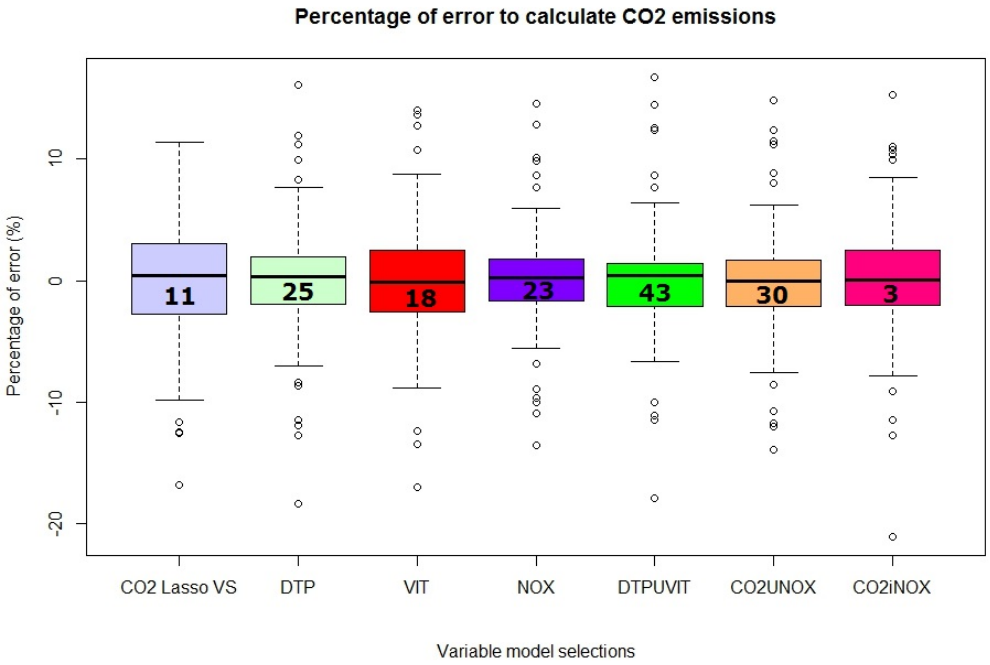



Fig. 3.53 – *Error distributions for models used to estimate the network $CO_2$ emissions and the number of links selected for each model.*

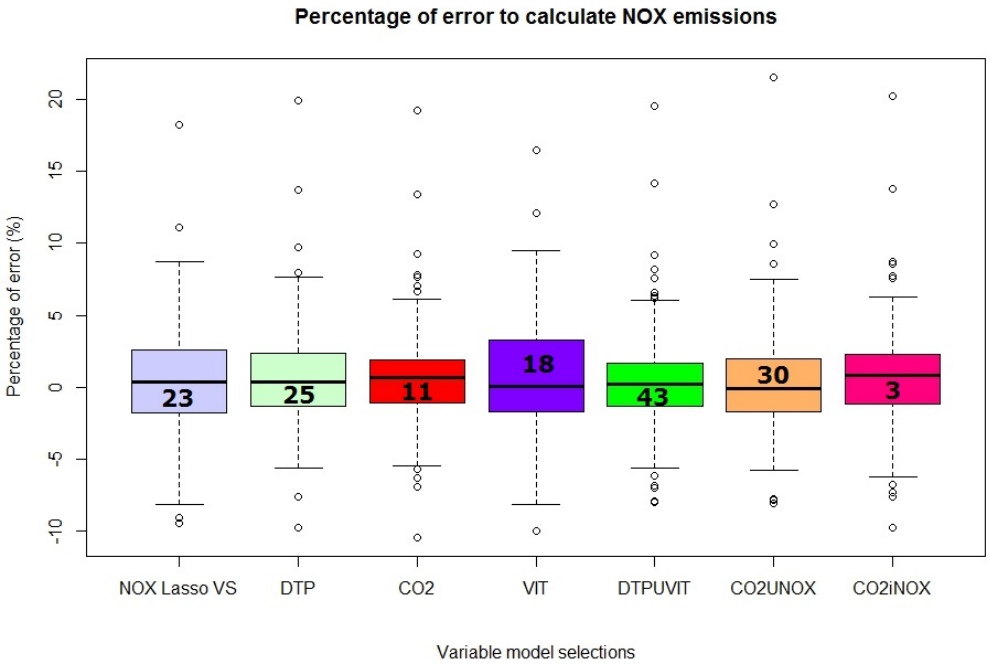



Fig. 3.54 – *Error distributions for models used to estimate the network $NO_x$ emissions and the number of links selected for each model.*

Considering all the models used to estimate the network traveled distance, $CO_2 \cup NO_x$ was the model with the lowest score meaning that it evaluates the network traveled distance better than the other models studied in this section. Two models have $BIC$ scores close to that of the lowest score, namely the $NO_x$ model followed by the $CO_2$ model, followed by the lambda model.

For the spatial mean speed, the lambda model is the best according to the $BIC$ scores. This score is followed by the 1 SE lambda and $DTP \cup VIT$. The other models had high scores compared to

| VARIABLES → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| MODELS ↓ | Validation set | Validation set | Validation set | Validation set |
| Reference | 3931 | 85 | 5652 | 4076 |
| Lasso Lambda | 3815 | -31 | 5554 | 3974 |
| Lasso 1SE | 3811 | -22 | 5563 | 3980 |
| DTP | - | 36 | 5557 | 3970 |
| VIT | 3850 | - | 5571 | 4000 |
| $CO_2$ | 3805 | 51 | - | 3972 |
| $NO_x$ | 3798 | 10 | 5521 | - |
| DTPuVIT | 3840 | -20 | 5553 | 3974 |
| DTP∩VIT | - | - | - | - |
| $CO_2$u$NO_x$ | 3797 | 6 | 5534 | 3985 |
| $CO_2$∩$NO_x$ | 3823 | 90 | 5572 | 3992 |

Tab. 3.18 – *Bayesian information criterion for all models studied in the dynamic/dynamic dataset.*

these three. Considering the $CO_2$ emissions, the best model was the $NO_x$ with a score equal to 5521 followed by $CO_2 \cup NO_x$ with 5534 points. For the $NO_x$ emissions, the best score was observed for the traveled distance (DTP). This was followed by three models with similar scores: the $CO_2$ model with a score equal to 3972, followed by the $DTP \cup VIT$ and LASSO lambda models, both with 3974 points on the $BIC$ scale.

**Conclusion on the dynamic/dynamic dataset**

The dynamic/dynamic dataset gives as results the daily traffic and emissions values. The predictors of this dataset are the time periods of each link and while the observations are the 15 minute variable values which correspond to the links and time periods. The matrices are compared with a vector that represents the network daily values (all the links and time periods grouped together) for each simulation. The purpose of the dataset is to identify in each link which time periods are really relevant.

Compared to the other datasets presented, the dynamic/dynamic dataset had the lowest sampling rates, which means less than 1% of predictors were used to fit a model used to estimate the network variables. Taking into account all the variables, increasing the size of the model does not decrease the errors, which means that all the models have similar error distributions.

The correlations between all the variables have a strong impact on the selection. For all variables, it was observed that the time periods that had strong correlations between them were also those that were selected most. If the traffic state is taken into account for the linear variables such as traveled distance, $CO_2$ and $NO_x$ emissions, the percentage of time periods in free-flow state is higher than congested time periods. As spatial mean speed is inversely correlated with all the variables, the congested time periods are selected most.

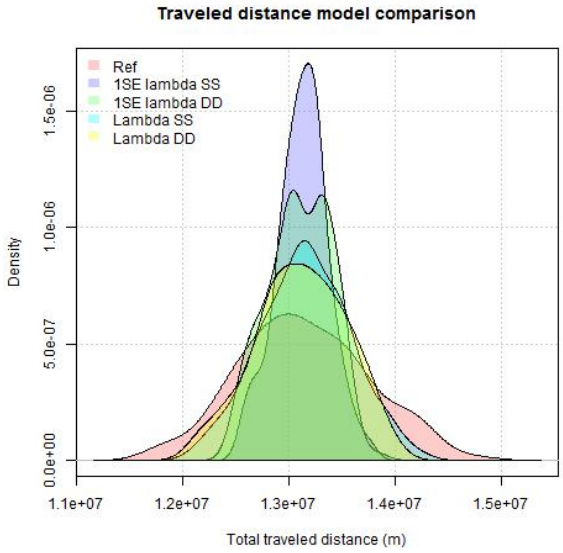
In comparison to the other models, the $NO_x$ model can estimate other variables and obtain good BIC scores. As for the $NO_x$ variable, the DTP or $CO_2$ models can be used.

### 3.3.4 Comparison between datasets

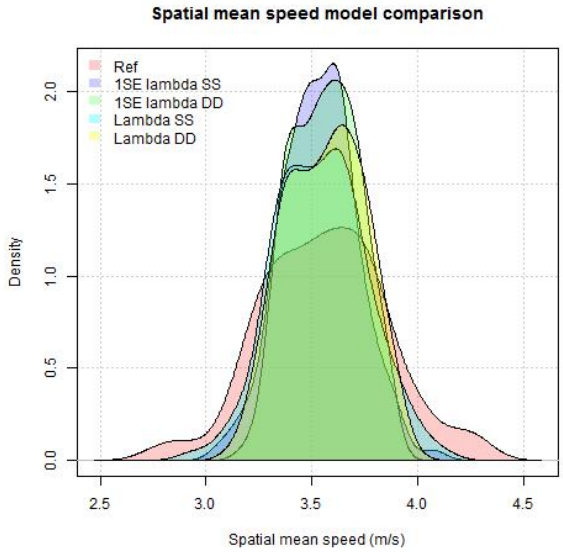
#### static/static versus dynamic/dynamic

The static dataset and dynamic dataset give the same type of results: daily traffic and emission values. The difference between them are: (i) the predictors of the static dataset are the links of the network while its observations are the daily values of each link according to the variables under study; (ii) the predictors of the dynamic dataset are the time periods of each link while its observations are 15 minute variable values that correspond to the link and period of time. Both sets are compared to a vector that represents the network daily values (all links and time periods grouped together) for each simulation. The aim of the first dataset is to identify the most relevant links in the network using daily values. In the second, the aim is to identify which time periods are really relevant in each link. This method can help to identify where it is possible to install on-road sensors, in reality, to estimate the network variables. The first analysis leads us to conclude that the model that selects daily traveled distance is the best as it selects only 3% of the network links with an average absolute error lower than 6%.

It is also important to consider that the static/static dataset can estimate network variables using fewer links (daily values) than those considered in the dynamic/dynamic dataset, although the data correspond to 15 minute aggregations. All these conclusions can be seen in 3.55.




(a) Traveled distance densities.




(b) Spatial mean Speed densities.

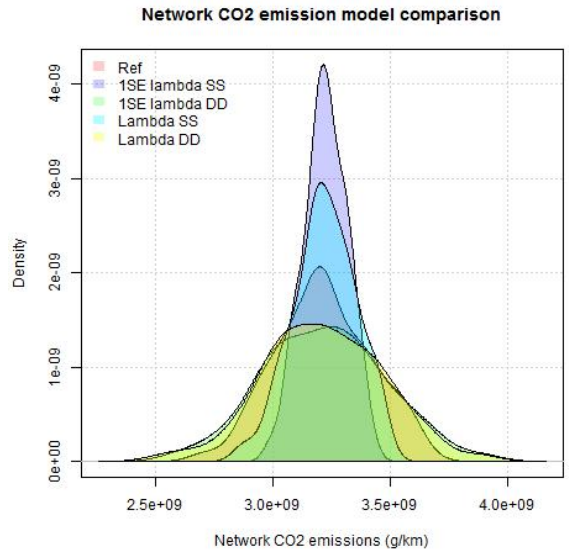
#### static/static versus static/dynamic

This comparison helps to answer the question: Is it possible to use models fitted for long time periods, such as daily ones, to estimate the variables within a short time range?

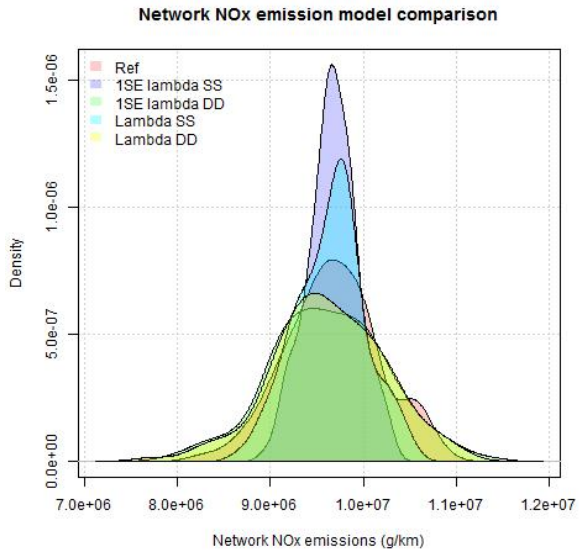
The models built with static/static datasets were applied to the same static/dynamic validation dataset by considering all the variables ($SS2SD$). For each variable, we compared their result values with their respective LASSO models and the reference values ($Y$).

The density distribution of variable values is presented in figure 3.56 to compare their results.

The resulting values from the models are quite different between the traveled distance, $CO_2$ and the $NO_x$ models. The static/static dataset models were fitted to daily values and when the same

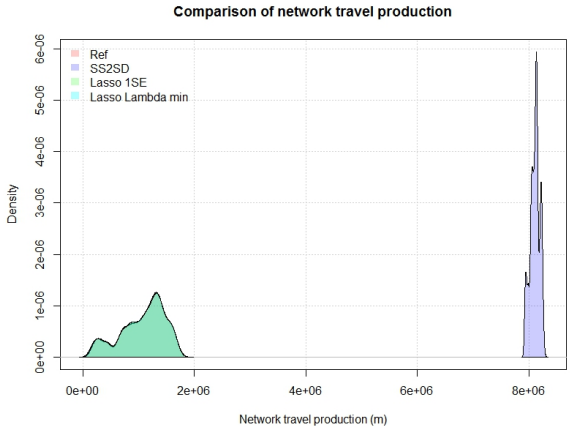



(c) $CO_2$ emission densities.

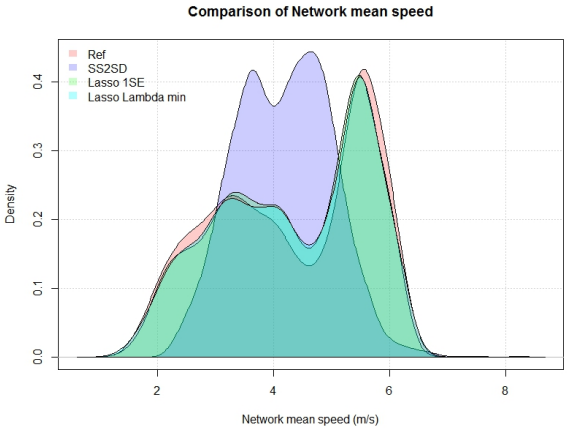



(d) $NO_x$ emission densities.

Fig. 3.55 – *Comparison between both optimized models with static/static and dynamic/dynamic datasets.*




(a) Traveled distance densities.




(b) Spatial mean speed densities.

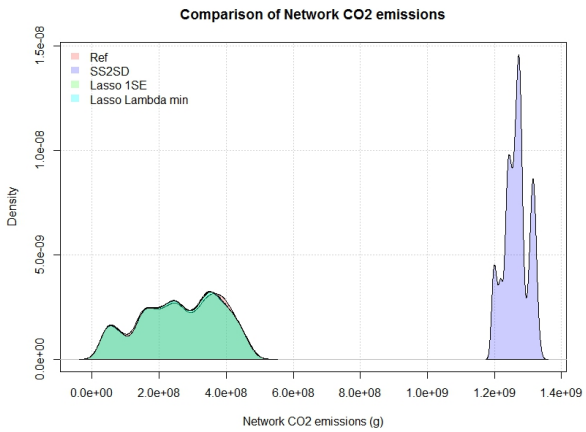
models were applied to the data representing a time period we obtained estimation errors over 500%. The difference of the temporal range between both datasets tends to overestimate the network values.

The error distribution for each time period was also analyzed using the static/static models (SS) applied to static/dynamic values (SD). Starting with traveled distance, their error distribution by time period is shown in figure 3.57 for each variable.

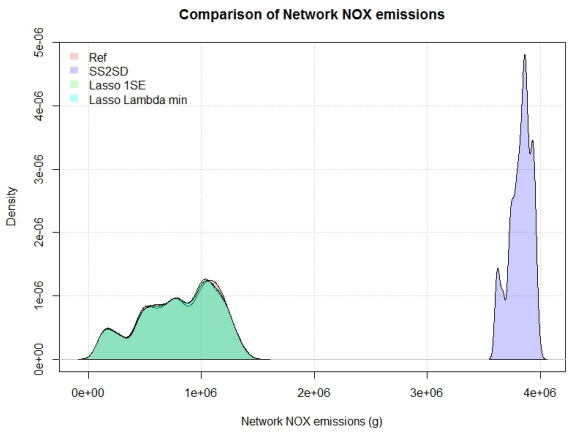
The percentage error is calculated using the difference between the reference values and the predicted ones from the SS2SD model (SS2SD = static/static model applied to static/dynamic data). On examining each period of time, it is possible to conclude that the error tends to be reduced for congested periods, but in general, the errors continue to be high.

For spatial mean speed, the results are better than those presented for traveled distance. Figure 3.58 shows the error distributions of spatial mean speed by time period.

In certain cases, when considering the error threshold to determine the spatial mean speed the errors are smaller for the free-flow states defined in table 3.18, than for the congested periods. The




(c) $CO_2$ emission densities.




(d) $NO_x$ emission densities.

Fig. 3.56 – *Comparison between: the reference values, the LASSO models and the application of static/static models to static/dynamic datasets for each variable.*

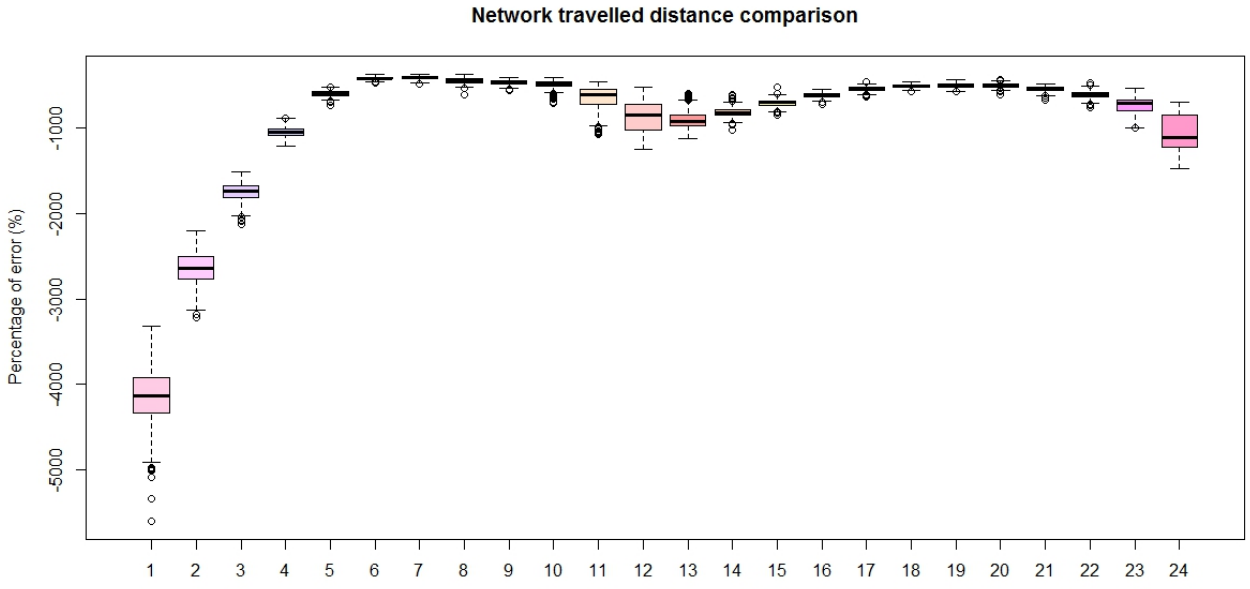



Fig. 3.57 – *The error distribution by time period for network traveled distance using the SS model applied in SD data.*

errors are smaller in comparison to the traveled distance variables.

Considering the pollutant emissions, both presented the same trend as for the traveled distance variable. They have large errors that prevent estimating the network variable values. Their results by time period are presented in figure E.1 for $CO_2$ emissions and in figure E.2 for $NO_x$.

In general, the LASSO models that were built using link data with time periods grouped (range of daily values) and applied to the same data ranging from 6 hours to 15 minutes, tended to overestimate the network variable values, leading us to conclude that it was not possible to use a model fitted to a daily time scale to estimate the variables in the shortest time period.

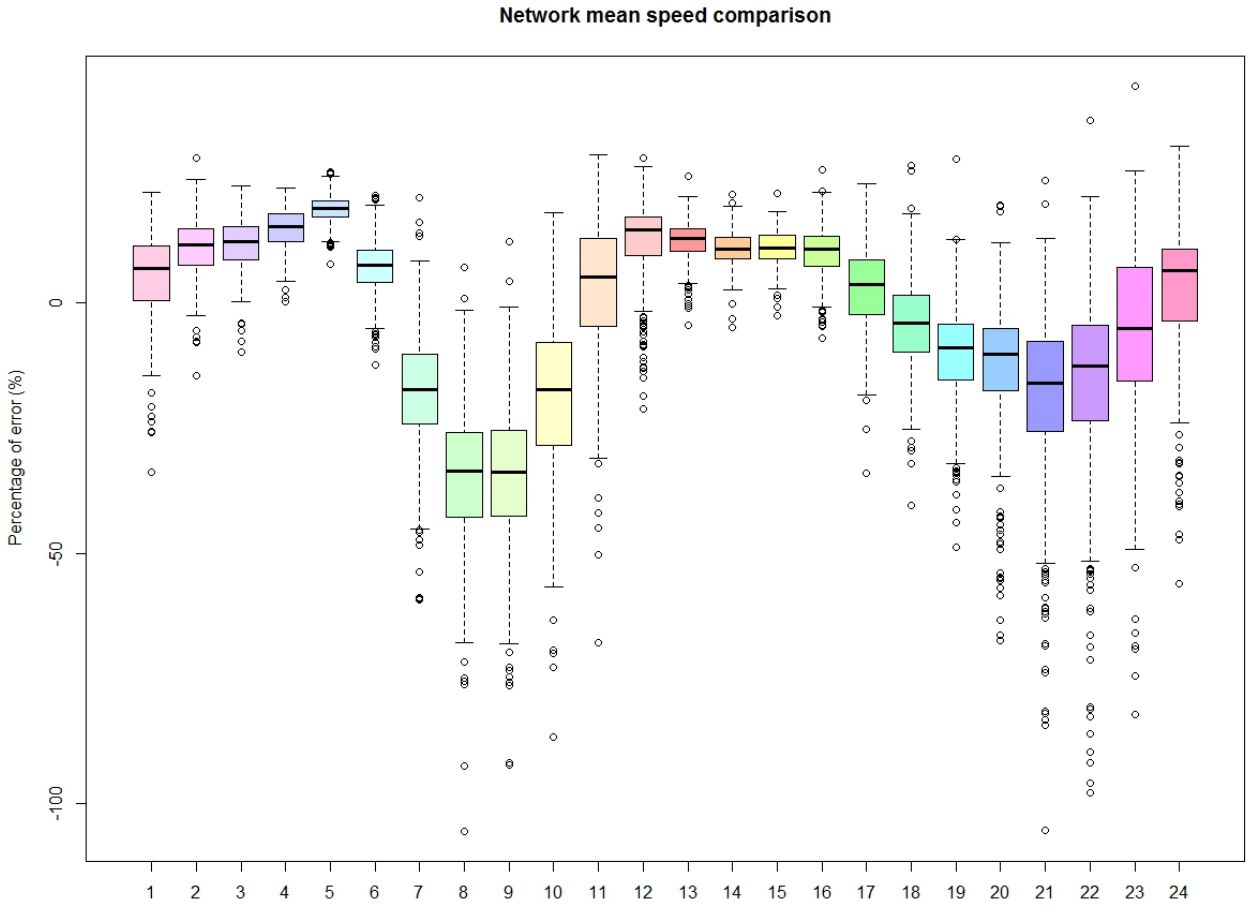



Fig. 3.58 – *The error distribution by time period for network mean speed using the SS model applied to SD data.*

### static/dynamic versus static/static

The same question posed in the analysis of temporal influence in the previous section is addressed here; but inversely. Can a model built with data described in the time range of 15-minutes be used with data that were gathered over a longer time (daily values). In figure 3.59 are shown the densities value comparison for each variable. The static/dynamic model (SD) applied to the static/static data (SS) was named $SD2SS$.

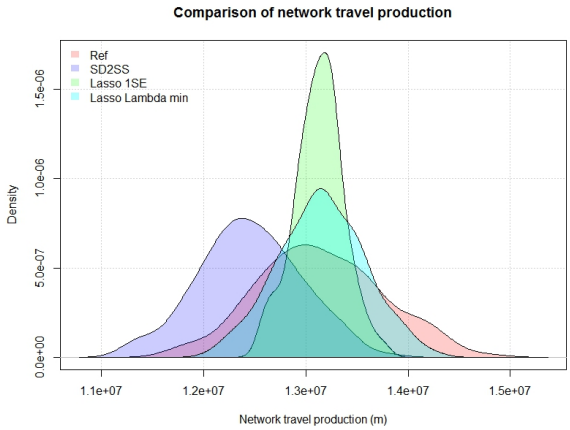
The $SD2SS$ results tend to overestimate all the variables (traveled distance, spatial mean speed, $CO_2$ and $NO_x$ emissions) when compared with their respective LASSO results and their reference values. Figure 3.60 show the error distributions for each variable.

The error distributions are calculated as:

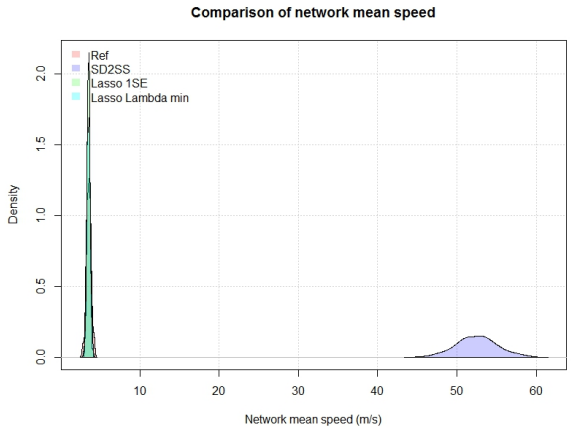
$$Error = \frac{Yreference - SD2SSvariable}{Yreference} * 100 \tag{3.8}$$

The model fitted with the 15 minute time period and then applied to daily values tended to overestimate the traveled distance variable by from 2.5% to 8%, as shown in 3.59 (a) and 3.60 (a). The network values are overestimated for the pollutant emissions between a little more than 1% and 10%. Finally, the variables are overestimated for mean speed with values that cannot be considered.

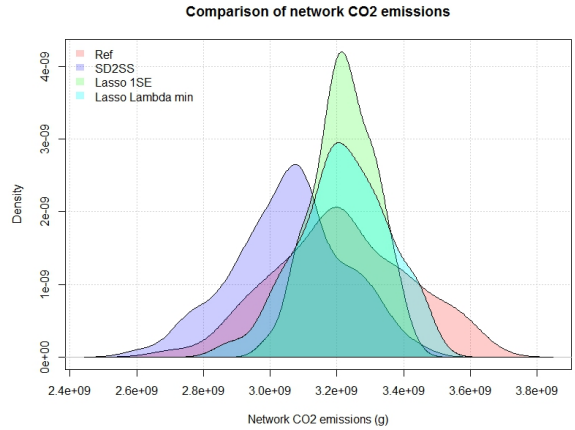
Using models fitted to data that represent the shortest time range to other data that represent the



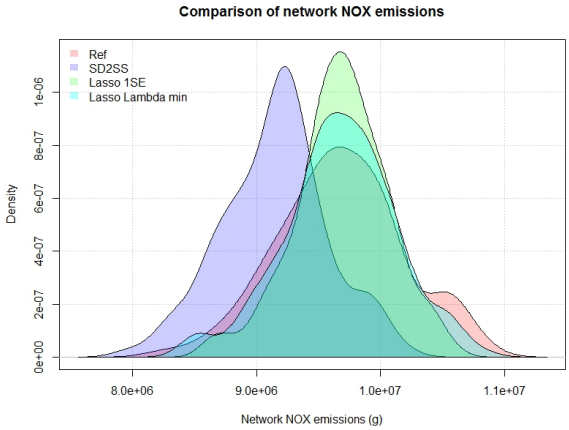

(a) Traveled distances densities.




(b) Spatial mean speed densities.

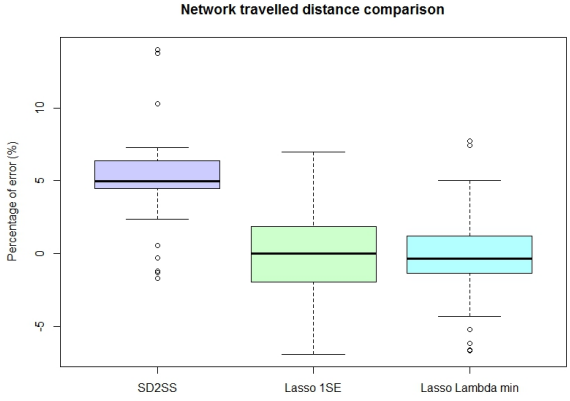



(c) $CO_2$ emission densities.

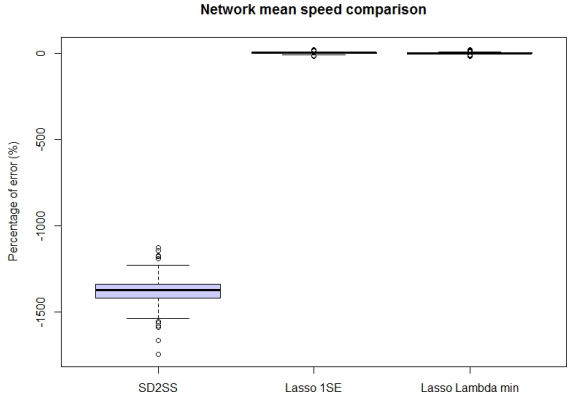



(d) $NO_x$ emission densities.

Fig. 3.59 – *Comparison between: the reference values, the LASSO models and the application of dynamic/static models in static/static datasets for each variable.*

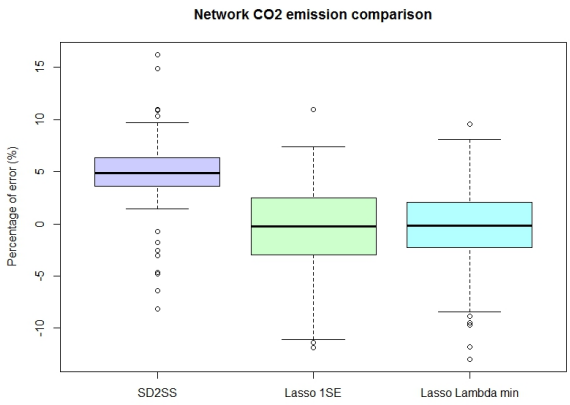



(a) Traveled distance error distributions.

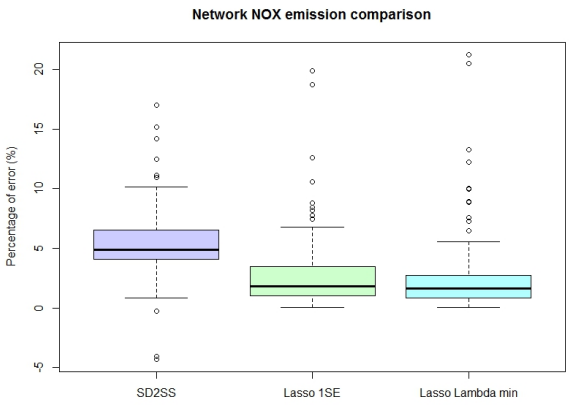



(b) Spatial mean speed error distributions.

daily values, leads to estimating variables with errors around +1% and +10% only for pollutants and traveled distances. The spatial mean speed cannot be estimated using this method.




(c) $CO_2$ emission error distributions.




(d) $NO_x$ emission error distributions.

Fig. 3.60 – *Error distribution calculated using the difference between reference values and the values from the application of dynamic/static models in static/static datasets for each variable.*

## 3.4 Conclusion of the chapter

This chapter described the study of the selection method known as LASSO using simulated data from the city of Paris. The main goal was to estimate the traffic and pollutant emissions on various spatial-temporal scales using only a set of links.

Three datasets were proposed to study the influence of the spatial and temporal scales on link selection: (i) the aim of the first, called the static/static dataset, was to estimate the daily values at the network level from the daily values at the link scale; (ii) the aim of the static/dynamic dataset was to estimate network values within a 15 minute time range using the same temporal scale at the links; and finally, (iii) the dynamic/dynamic dataset was built to identify the most relevant time periods from daily data to estimate the daily network values.

The first dataset highlighted the links that can be used by policymakers to decide where to implement sensors on network links in order to estimate daily values. The second dataset can be used to forecast traffic and emissions in the network for real-time tracking, as taking into account congestion gives a better estimation of traffic and emission values. The purpose of the last dataset was to study the spatial-temporal correlation, selecting not only links but also the most relevant time periods.

The LASSO method was chosen due to its capacity to: (i) estimate large numbers of predictors as a function of a limited number of available observations; (ii) correlate predictors, and (iii) carry out the bias-variance trade-off. The method was applied to all the variables and all the datasets. Considering the dataset (i), it was possible to estimate the network values using less than 5% of the links equipping the network with estimation errors within $\pm 5\%$ and $\pm 10\%$. For the static/dynamic dataset, more links were selected and they represented around 30% of the network links with an error distribution lower than $\pm 10\%$.

LASSO provides a large number of possible models with different model sizes and levels of error prediction. Among these models, the 1SE lambda was highlighted first as it provided a good compromise between model size and estimation error. A study of model size and estimation error was then conducted to examine how the estimation error evolved in models with more or fewer selected links in comparison to the 1SE lambda model. The 1SE lambda model remained the best compromise for (i) and (ii), while the lambda model was the best choice for (ii).

The merger between traffic variables and between pollutant emissions was proposed to observe whether combined selection could increase the precision of the estimation by reducing the errors. Similarly, the aim of the intersections was to reduce the number of selected links, taking into account the links in common between traffic variables and emission variables, and to observe how the estimation error was affected. The intersections presented a better compromise between the number of links needed and the estimation error in all cases.

After concluding that many possible sets of links could be used to estimate the network values and also that the links could be interchanged with each them without greatly affecting the estimation, we then explored whether only one set of links among all those studied could be used to estimate all the variables of the network.

Also, the influence of assignment on the selection was considered. The aim was to study the possibility of using only the morning or afternoon data to estimate the network values and vice-versa.

To end the study, the models fitted to a temporal range were able to estimate the network values using either a larger or smaller quantity of temporal data.

To qualify the effectiveness of the LASSO method for environmental assessment, these results must be confirmed with other data sets in different networks and compared to other selection methods. However, the daily emission value was often not sufficient when focusing on the exposure of the population to pollutants.

# 4

# Model selection from other statistical methods

Statistical techniques are commonly used in many aspects of traffic modeling and prediction. Some methods demand considerable computation effort regarding the large number of inputs and their variations. Other techniques can reduce computation time and effort and improve convergence to a representative result. This was the case of LASSO in the previous chapter. Other methods can be used for highly variable data and are based on the probability of occurrence of corresponding values to reduce their variance (Ang and Tang, 2007). In this chapter, emphasis is placed on simple statistical methods that have similar characteristics, such as linear regression. Here, the same data and links as in the previous chapter are subjected to other methods in order to compare them. Four methods were chosen: (i) random sampling selection; (ii) ranking links from the most pollutant to the least; (iii) stepwise selection which proposes its own model and optimal selection process; and (iv) a network partitioning method that clusters the network before selecting the links. For all these methods, the aim is to estimate daily emissions as well as 15 minute network pollutant values.

Method (i) is basic. The methodology used is the same as that in a lottery in which samples are selected randomly. The second method is based on ranking. The links are ranked by the amount of pollutant emissions they emit. The most pollutant ones are chosen to estimate the network emissions based on regression. Method (iii) uses multiple regression in which only the members of the population that improve the model's accuracy are selected. Like LASSO, the stepwise method selects the most relevant links. The last method, represented by (iv), first considers partitioning the network into clusters and then using the random sampling method applied to each cluster to finally estimate network values based on linear regression. The first two sampling techniques proposed use simple linear regression to build models and estimate the network values. Regarding methods (iii) and (iv), the aim is to compare other smart techniques to obtain representative samples of the network and possibly reduce the rate of errors on the estimations. The comparison between these models may provide useful information on the varying complexities and efficiencies of the sampling models used.

For this chapter only two datasets were selected from the last chapter. The first, called static/static, aims at estimating the daily network emissions from a local daily emission at link level. The purpose of the second, called static/dynamic, is the same as that of the static/static dataset but with a refined temporal level. Thus the 15 minutes network emissions will be estimated using the 15 minute local emissions at the link level. All the methods were applied to both datasets. Their results were compared with those obtained with LASSO and with each other.

The chapter ends with a description of a study using the location of real sensors installed in a district of Paris. Using their locations and the simulated values of emissions in these locations, an estimation of the network emissions is conducted with linear regression applied to two temporal ranges:

daily and 15 minute estimations. The results are analyzed and compared. The aim is to quantify the errors using the real location of sensors when estimating emissions at the network level.

## 4.1 Random Selection

The previous chapter showed that the network variables can be estimated with reasonable error using a simple linear regression. It also showed that many different combinations of links of the $6^{th}$ district of Paris can be interchanged without affecting the quality of the estimation. With this in mind, the possibility of using a naive method can be considered in order to compare it with the previous sampling technique.

Considering that we have the necessary data for all the members of our population (i.e. links), the basic random sampling technique can be used. The aim is to sample links randomly considering the entire network (i.e. 230 links). In order to make the LASSO and random methods comparable, the random sampling uses the same selection rate as the LASSO method. Thus, all the links can be selected with equal probability at any step of the sampling.

The previous chapter also showed that the spatial mean speed was the hardest variable to estimate in all the datasets because of its non-linear behavior. In contrast, traveled distance and pollutant emissions follow a linear pattern. Pollutant emissions follow the same behavior as traveled distance. Only the pollutant emissions, $CO_2$ and $NO_x$, are considered in this chapter.

Two random sampling methodologies were studied: the first one considers the entire network to build a sample at the same selection rate as LASSO did for each pollutant and dataset using the lottery method; and the second considers the LASSO selection rate distributed proportionally between cluster zones defined by the snake method described in (Ji and Geroliminis, 2012a). The purpose of the clustering will be explained in this section. All the results are compared with the main results of LASSO (called the 1SE lambda model) for each variable. The results will be presented for each dataset separately.

### 4.1.1 Static/static dataset

#### Lottery method: Random selection at the network scale

In the previous chapter, the 1SE lambda model built using the LASSO method selected the 11 most relevant links of the network to estimate the daily $CO_2$ and $NO_x$ emission values at the network scale. The selected links of both pollutants are identical in 72.7% of cases. The same quantity of links in the whole network was selected randomly to estimate the daily emissions. Each link has an ID number ranging from 1 to 230 for identification. The 11 ID numbers were generated randomly 30 times to take into account the diversity of the possible combination between the links of the network. The random methodology was applied for each pollutant separately. Although linear regression was used for the observation values of these random selected links, it does not guarantee that the random samples can represent the population perfectly. To evaluate the emission estimation based on the model built, the relative errors are calculated and compared between all the random draws.

Using the same training and validation sets of observations as used in LASSO, a simple linear regression was applied to the training set of each random draw. Then, the error distribution of each one was calculated after applying the linear model fitted to the validation set.

Figure 4.1 shows the error distributions for each random draw of $CO_2$ emissions.

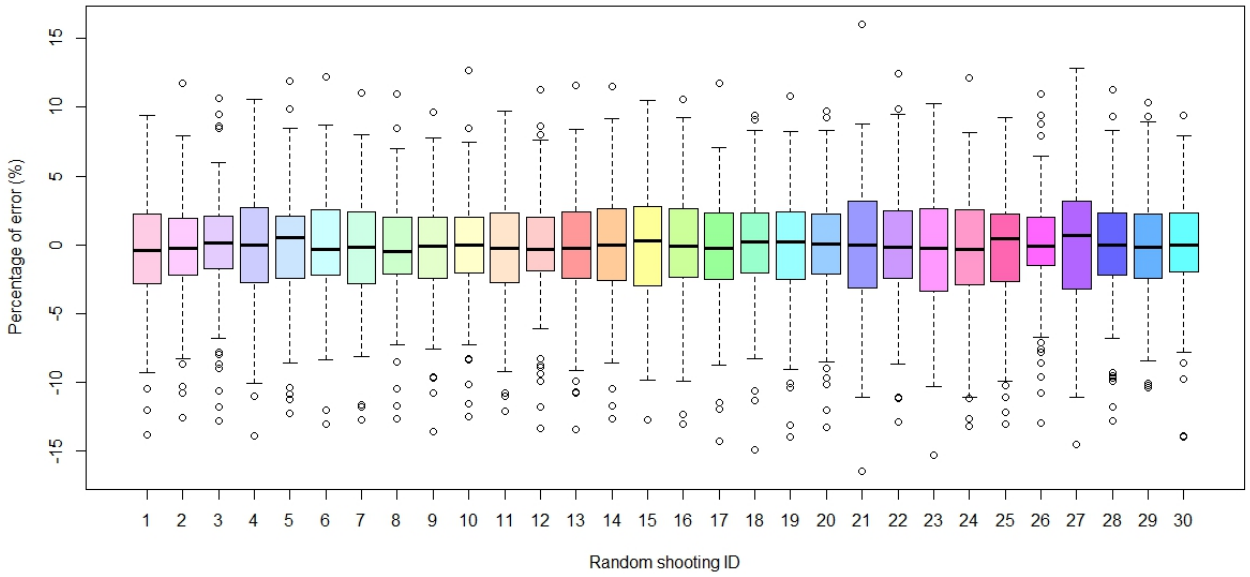



Fig. 4.1 – *The error distributions of the daily $CO_2$ emission estimations of each random sampling.*

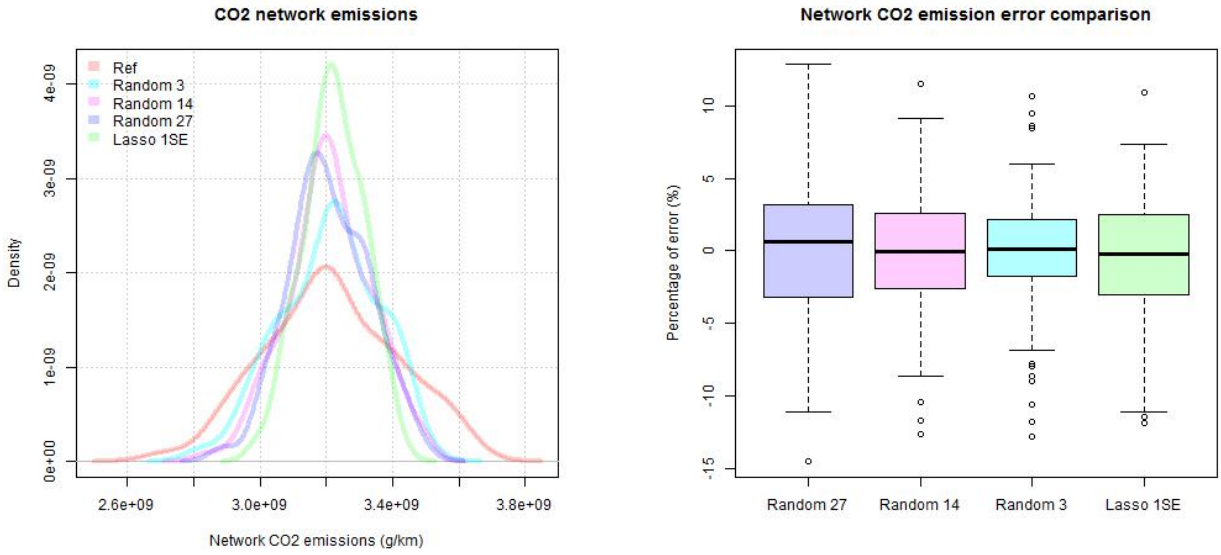
As can be seen in figure 4.1, most of the random selections have errors between -10% and 10%. Only a few have an error distribution a little higher than ±10%. Also, 50% of the data distributed in each boxplot have a percentage error lower than ±5% and most of them have a centralized error distribution with a median value around 0%. All the plots also present a minimum number of outliers, with up to ±15% percentage error.

As presented in chapter 2, the network has certain links through which no passenger cars passed during the simulations, and consequently these links had no associated emission (i.e. zero values). Among the 230 links, 23 are without car flows. The LASSO method does not consider these links automatically in the selection, contrary to the methodology defined for the random sampling, which considers the entire network without exception. This type of link can be selected by the lottery method, which means that some models have fewer predictors than the selection rate defined by LASSO due to the null links.

The 1SE lambda model for the $CO_2$ emissions presented an error distribution between -10% and 8%, and most of the random draws showed similar error distributions. As shown in figure 4.1, the majority of the random draws have the same or fewer dispersed errors. Of all the random draws, only 7 have the 11 selected links with car flows, the other ones selected between 1 and 3 links with no flow. Taking the latter into account, the majority of the random draws present model sizes smaller than that defined by the 1SE lambda model built using the LASSO method. For example, the random sample ID number 3 is that with the least scattered errors compared to the others and it has only 10 links (non-null) in the model. The worst case of all the samples is ID 27 which presents the largest errors, lower than ±15%, with only 8 non-null selected links inside the model. ID 14 is an intermediate case with 11 non-zero links.

This leads to the conclusion that some outputs of the optimal selection performed by LASSO can be switched to another one without a significant increase of errors. The comparison between the random selections and the LASSO estimation are presented in figure 4.2.

Figure 4.2 shows the density distribution of the values predicted by the models in ($a$) and the




(a) Daily $CO_2$ emission densities.

(b) Error distribution of daily $CO_2$ emission estimations.

Fig. 4.2 – *Comparison of the $CO_2$ estimations from random selection and the LASSO model.*

associated errors of the predicted values in (*b*) when compared with the reference values. As discussed previously, the estimated values of the daily $CO_2$ emissions from the random models are similar to those estimated by LASSO. The three random cases presented in the figure represent the worst, the intermediate and the best cases, respectively, of all the random draws. All of them have selected links in common with those selected by LASSO. As noted, the level of estimation errors is of the same order as for the LASSO model.

The methodology applied to the $CO_2$ emissions was also applied to the daily emissions of $NO_x$. The 30 new ID random draws were performed to select 11 links of the network. The 1SE lambda model estimated the daily $NO_X$ emissions with error distributions lower than ±8%. The error distribution for each random sample is shown in 4.3.

Comparing the samples of both pollutants, the $NO_x$ emissions had fewer dispersed errors than the $CO_2$ emissions. They presented percentage errors between -10% and 10% without considering the outliers. Of all the samples, only 8 had a full-size model, which means no null links were selected, 13 out of 30 selected only 1 null link, 6 out of 30 selected 2 links and finally 3 out of 30 selected 3 links without car flows. In figure 4.4, the sample with the least dispersed errors is ID 15 with a percentage error lower than ±5% and a model size equal to 10. The worst case is ID number 2, with a little less than ±10% error and a model equal to 8 links. ID 20 is an intermediate case compared to the previous ones, with a model size equal to 9 links and the same error distribution as the 1SE lambda model built using the LASSO method.

Figure 4.4 shows in (*a*) the densities of the predicted values from the three selected cases of random models, the LASSO model and the reference values; and in (*b*) the associated errors of these models. As in the $CO_2$ emissions, the density of the estimated emission values of the worst, intermediate and best models from the random method is similar to the density of those estimated by LASSO. Also, the errors of predicted values are in the same range as those from LASSO, on average, and in some cases they can be better with a smaller model.

The conclusions made for $CO_2$ are also applied to $NO_x$ emissions. The random sampling method,

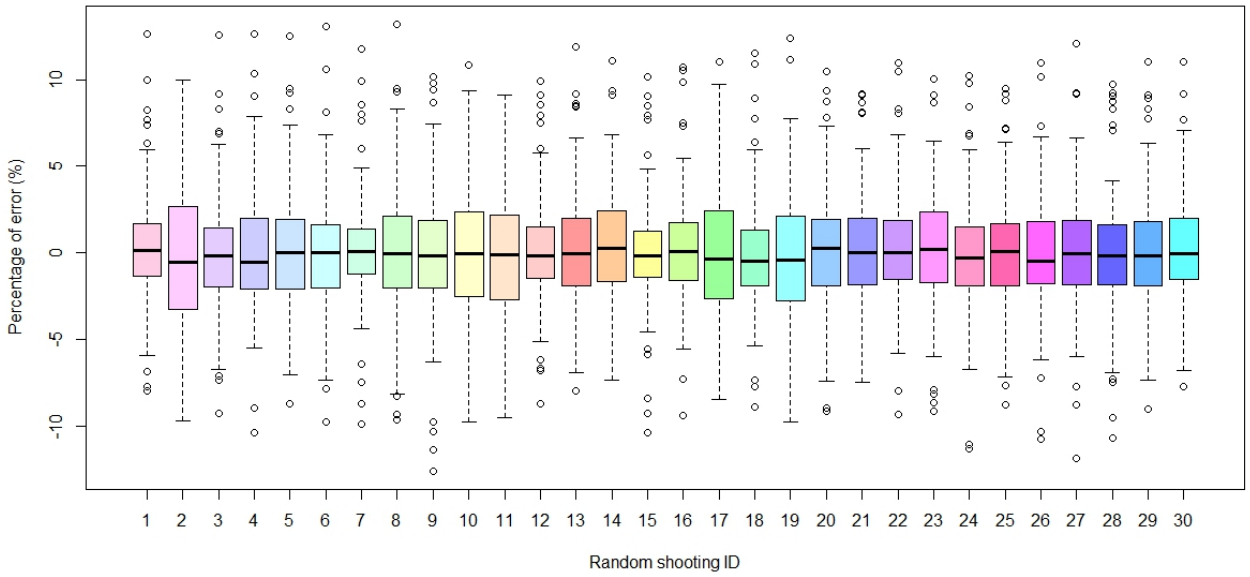
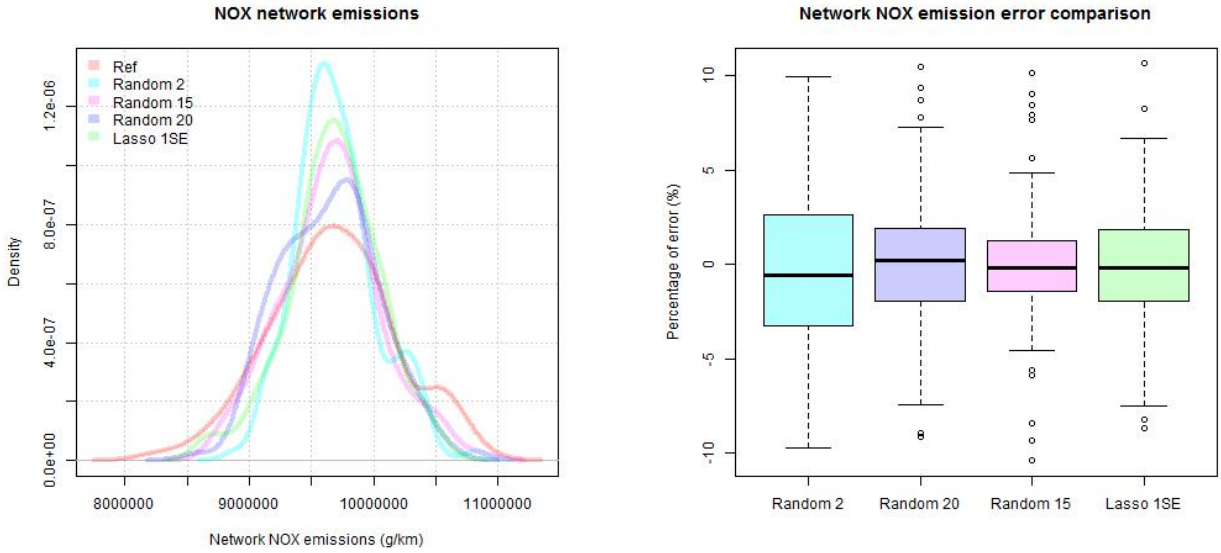



Fig. 4.3 – *The error distribution of the daily $NO_x$ emission estimation for each sample.*




(a) Daily $NO_x$ emission densities.

(b) Error distribution of daily $NO_x$ emission estimations.

Fig. 4.4 – *Comparison of the $NO_x$ estimations from the random selection and the LASSO model.*

at the rate defined by the sampling carried out in the previous chapter using the LASSO method, presents a similar error on estimations compared to LASSO with different sets of links. Considering that for LASSO the whole network was used as input to apply the sampling, and that it automatically excludes the links with no car flow, the same network was considered using the randomized sampling. The difference between both methods is the possibility that a null link can be selected, thus some models can be smaller when compared to those obtained with the 1SE lambda model.

Even when considering smaller models and various combinations of links for both pollutants, the randomized sampling obtained estimations with almost the same degree of error as the smart method,

i.e. LASSO. For this neighborhood of Paris with its traffic characteristics, randomized sampling and a linear regression can be employed to estimate daily network emissions using day-to-day emissions at link level. Furthermore, the method needed less than 5% of the total links and obtained estimations with errors around ±5% for the best cases and less than ±15% considering the worst ones. The estimations are in the same range of errors as that of the LASSO method demonstrated in the previous chapter.

**Random selection in a cluster**

In this section the same lottery method of sampling was applied to clusters. The clusters were defined by a partitioning method based on similarities between links. The random selection was carried out proportionally as a function of the size of the clusters. Considering the network's spatial-temporal heterogeneity, clustering helps to define compact and homogeneous zones, that is to say a group of links related with others around them that have similar traffic or emission characteristics.

A large number of clustering methods have been designed specifically for traffic networks. Here, we use that proposed by (Ji and Geroliminis, 2012a) and (Saeedmanesh and Geroliminis, 2015), *i.e.* the *NCut* based on *snake* similarities. *NCut* is a partitioning method that can divide a graph into spatial zones (Shi and Malik, 2000). The literature comprises numerous proposals to adapt it to traffic networks, by using a method to calculate a similarity matrix designed to split up a network optimally.

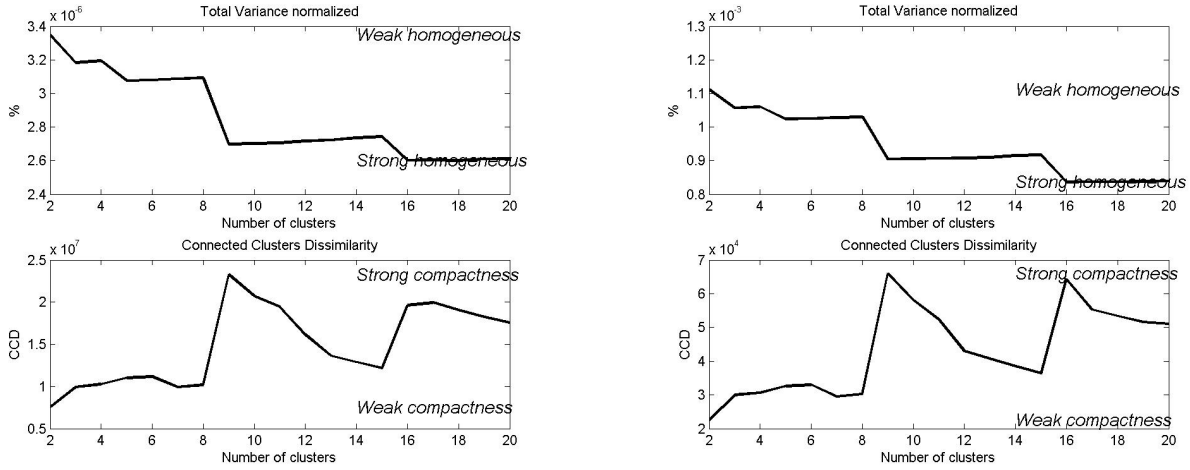
Basically, the principle of *snakes* starts with a link of the network and then it adds the neighboring links one by one when similar characteristics are found, in our case pollutant emission levels. The *snake* is expanded only with links that are considered adjacent and always starts with a single link of the network before exploring the entire network. In addition, the size of the *snake* depends on the variance, with the aim of minimizing it. For more details regarding the method applied and improved to correspond to a traffic network, see (Lopez et al., 2017).

Two main indicators were used to assess the goodness of the partitioning method used for the Paris network: total normalized variance and connected cluster dissimilarity. The former is an indicator of emission homogeneity based on the inter-cluster link emission variance. The second is an indicator of the dissimilarity between adjacent clusters. The figure 4.5 shows both indicators for the $CO_2$ and $NO_x$ emissions.

The shapes of the indicators are fairly similar for both pollutant emissions. The figures show the optimal number of zones for a cluster is 18, where the clusters have the strongest homogeneity and compactness. But considering that the network of Paris comprises 230 links, dividing them into 18 zones is unrealistic and cannot be considered for a real case. Thus, it was decided to partition the network into 4 and 6 zones for both pollutant emissions. This number of clusters was chosen to obtain almost the same level of dissimilarity, a slightly different level of homogeneity and to make it possible for the cluster to have different snake sizes. They will be the same for both the datasets studied.

We first present the results for $CO_2$ emissions. Figure 4.6 shows the 4 and 6 zones defined by the partitioning method.

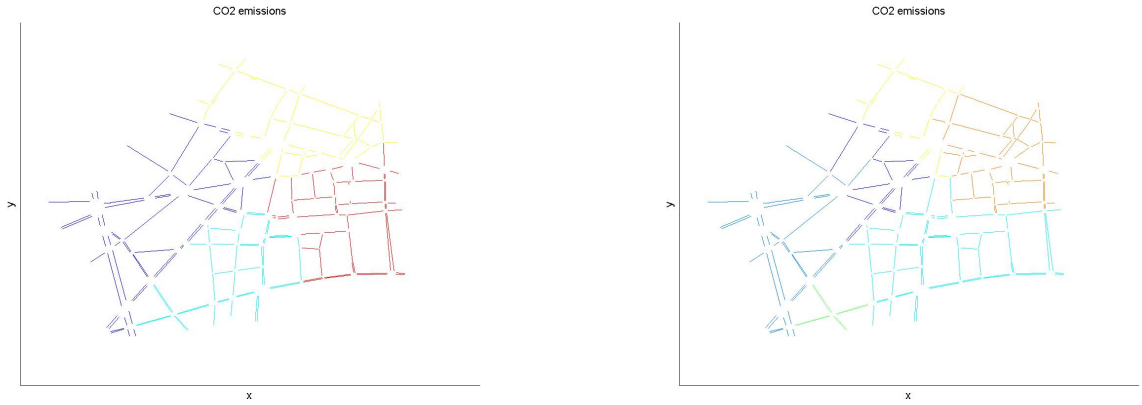
The clusters are identified by different colors and their sizes determine the proportionality of the total number of links in the random draws. Considering ($a$), their sizes varied between 20% and 30%.The clusters are described as follows: (i) the red cluster has 59 links which represent 25.7% of the total links; (ii) the yellow cluster has 55 links representing 23.9% of the network; (iii) the cyan cluster has 48 links representing 20.9% of the total; (iv) the blue cluster is the largest with 68 links representing 29.5% of the network. Taking into account that the LASSO selection gave 11 links (4.8%




(a) Indicators of $CO_2$ emission partitioning.

(b) Indicators of $NO_x$ emission partitioning.

Fig. 4.5 – *The total normalized variance and the connected cluster dissimilarity indicators of $CO_2$ and $NO_x$ missions.*




(a) The $CO_2$ emission partitioning in 4 clusters.

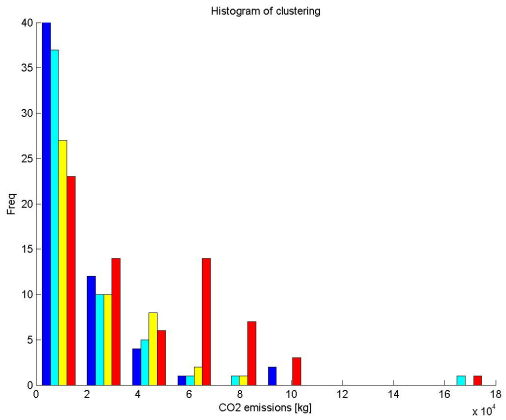
(b) The $CO_2$ emission partitioning in 6 clusters.

Fig. 4.6 – *The partitioning of the Paris network based on daily $CO_2$ emissions.*

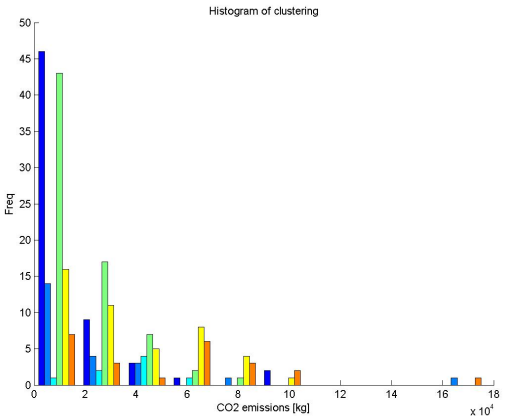
of the network), these 11 links will be selected proportionally as a function of the cluster's size, thus 3 links each for the red, yellow and blue clusters, and 2 links for the cyan cluster, totaling 11 links selected randomly.

Considering the second partition in 6 clusters shown in (*b*), the sizes are more variable compared to the previous partition and they correspond to between 3% and 31% of the network. The clusters are described as follows: (i) the orange cluster is composed of 61 links representing to 26.6% of the network; (ii) the yellow cluster has 23 links representing 10% of the network; (iii) the size of the green cluster is equal to 8 links (3.5%); (iv) the cyan cluster is the largest one with 70 links (30.4% of the network); (v) the light blue cluster has 45 links (19.5%); and, finally, (vi) the blue cluster has 23 links (10% of the network). The random selection into clusters is distributed as follows: only one link was selected randomly in clusters (ii), (iii) and (vi); 2 links were selected in cluster (v) and, finally, 3 links in cluster (iv).

Also observed is the contribution of each cluster to emissions in both cases. Figure 4.7 shows the clustering of $CO_2$ emissions.




(a) Histogram of $CO_2$ emission partitioning into 4 clusters.




(b) Histogram of $CO_2$ emission partitioning into 6 clusters.

Fig. 4.7 – *Histogram of clustering in the Paris network based on daily $CO_2$ emissions.*

The histograms show the distribution of links by range of $CO_2$ emissions. In $(a)$ the red and yellow clusters each have similar contributions of total emissions and correspond to 17% on average. The cyan cluster represents 18% and the blue cluster represents most of the $CO_2$ emissions with 48% of the total.

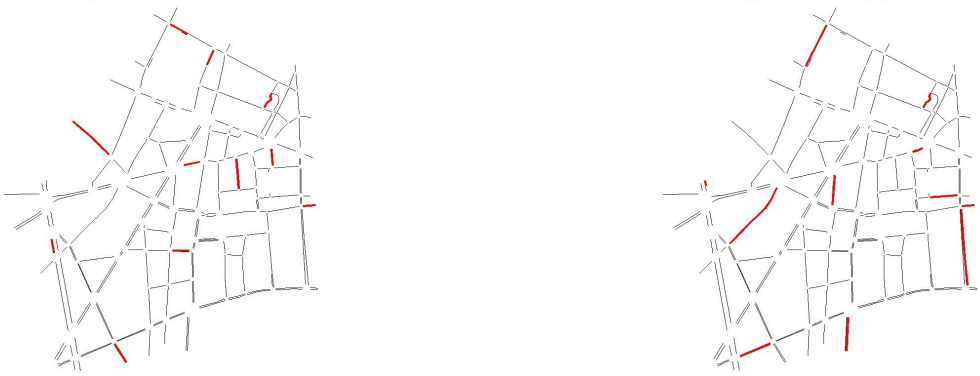
Considering the network divided into 6 zones, their contributions to the total emissions are: orange cluster 15%; yellow cluster 10%; green cluster 5%; cyan cluster 22%; light blue cluster 27%; blue cluster 20% of the total $CO_2$ emissions. Although the cyan cluster is composed of 70 links, the contribution of the total emission from these links is only 22%. Thus, it can be considered that the links of this cluster have low emission rates compared to the others.

After defining the clusters in the network, the random selection was performed for each case, as described before. The links selected for both partitions are shown in figure 4.8. As can be observed between both random selections distributed over the network, there are only 2 links in common and one of them is a link with no car flow (i.e. associated zero emission value). Considering the random selection in in $(a)$, only two of the 11 links are null-links, which means a model built with 9 predictors. There is only one null-link in $(b)$, giving a total of 10 predictors for the model. In both cases, the model built by linear regression will be smaller compared to the LASSO model used as the reference.

Following the random selection of links, a linear regression was applied to the same training set of observations as that defined in LASSO, and the models built were applied to the validation set to quantify the associated errors. The daily estimated values of $CO_2$ emissions are shown in figure 4.9.

When analyzing the density of the estimated daily $CO_2$ emission values, the values from both partition cases are closer to the density of the reference values than those estimated using the LASSO method, as the latter tends to estimate values around the mean. When the networks are divided into 4 clusters, the distribution has almost the same shape as that obtained with LASSO, while the estimated values from 6 clusters are more similar to the reference values. This leads to the conclusion that partitioning the network into homogeneous zones significantly increases the goodness of fit, with a corresponding reduction of errors.

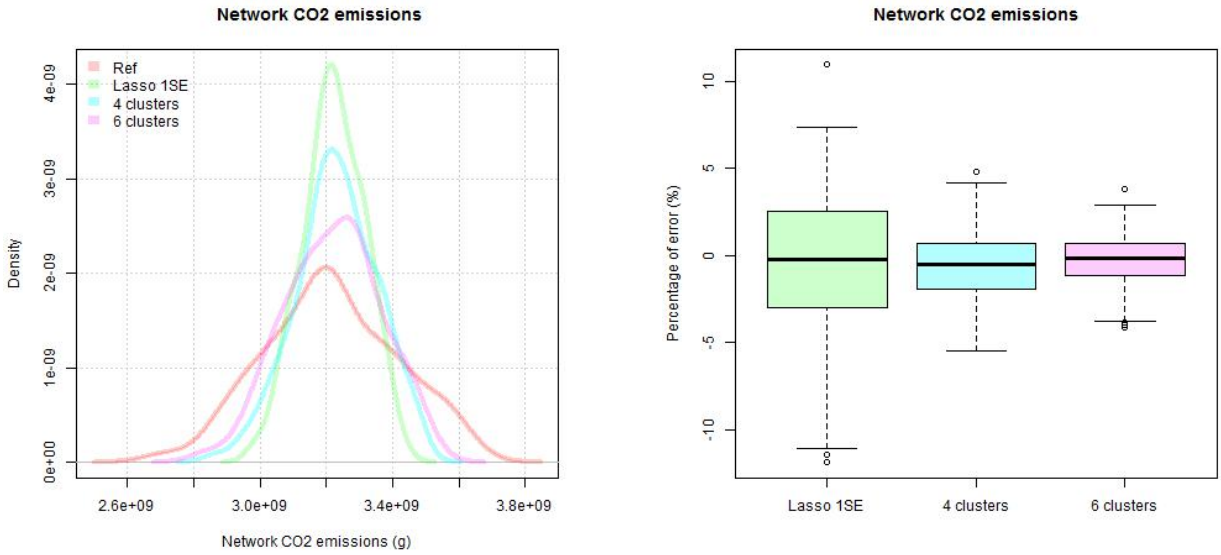
The percentage errors of the network divided into 4 zones vary from -5.5% to 4.8%, considering that 50% of the data analyzed have percentage errors from -1.9% and 0.7%. The errors have a centered distributed with a median value equal to -0.5%. Regarding partitioning into 6 clusters, the percentage



(a) The random links selected from the network divided into 4 clusters.

(b) The random links selected from the network divided into 6 clusters.

Fig. 4.8 – *The 11 links selected randomly in both cases of partitioning.*




(a) Density of estimated values of $CO_2$.

(b) Associated errors of daily estimated values of $CO_2$ emissions.

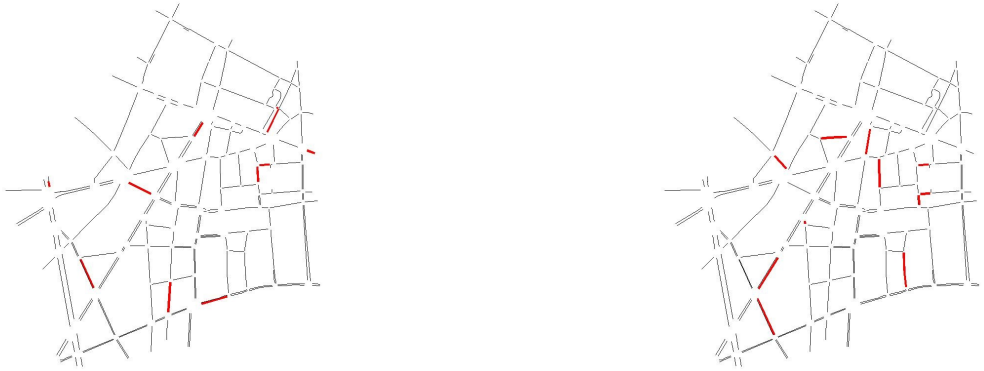
Fig. 4.9 – *Density and associated error for the estimated $CO_2$ emission values for the LASSO model and the models based on the random selection into clusters.*

errors are even smaller, from -3.8% to 2.9%. 50% of the estimated values have errors around -1.1% and 0.7%. In this case, the error distributions are more centered than the previous cluster partitioning, with a median value equal to -0.1%. The LASSO method has much higher percentage errors that can reach around ±10%, with a percentage error around ±3% for half of the data. Regarding the outliers presented, the 4 cluster errors have only one case and the 6 cluster errors have 4 outliers, which represent 0.8% and 3.0% of the data respectively. The LASSO model also presents lower rates of outliers but with larger errors than the partition method.

Therefore, taking into account that the partition method increases the goodness of the estimated values with fewer predictors than defined by the LASSO method, the partition method proved to be a better method for estimating network emissions. Its methodology allows defining regions with similar characteristics, which makes any link inside a cluster more representative than a random selection

over the network.

For the second pollutant emission, i.e. $NO_x$, the network was partitioned in the same way as for the $CO_2$ emissions, with respect to their cluster sizes, the contribution of their emissions as a function of the total, and the number of links selected in each cluster. The detailed explanation of the clusters considering the $NO_x$ emissions is presented in the appendix F.1. The results from the models after the clustering will be addressed here. Figure 4.10 shows the random selected links from the network partitioned into 4 and 6 clusters.



(a) The random selected links from the network divided into 4 clusters.

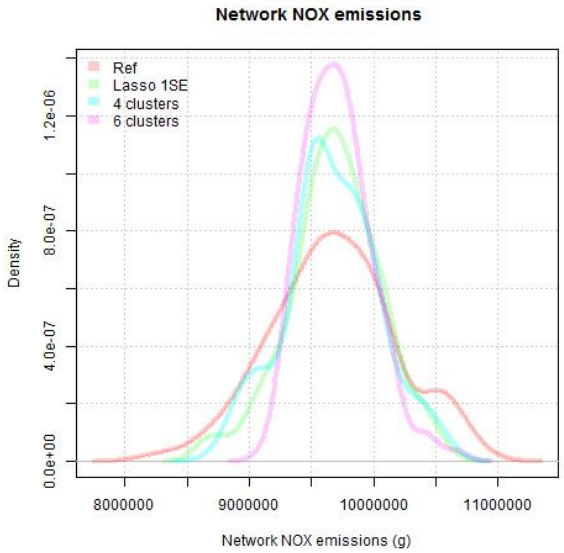
(b) The random selected links from the network divided into 6 clusters.

Fig. 4.10 – *The 11 links selected randomly in both cases of partitioning.*

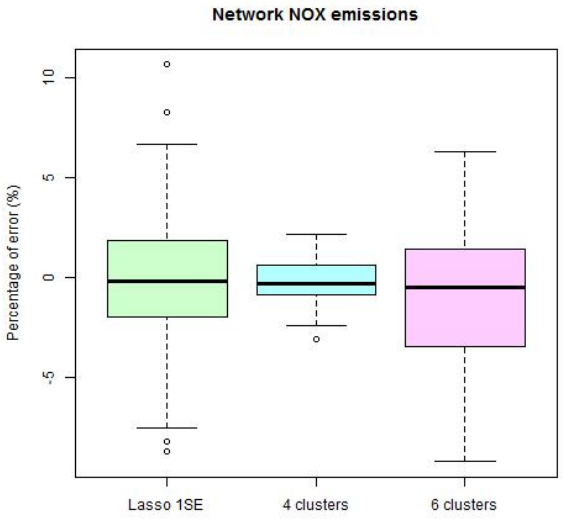
Both random draws share only one common link. The null-links were also selected in both sets of links, 2 in the 4 partition cluster and 3 in the 6 partition cluster. Consequently, the models were built with only 9 and 8 predictors, respectively. The linear regression function was applied to the training set of observations and the errors associated with the model were calculated using the validation set. The comparison of the estimated values and the associated errors from both partitions are shown in figure 4.11.

Although the $CO_2$ partitioning is identical to that of $NO_x$, the estimated values of the latter differ from those of the estimated $CO_2$ values. The model fitted to the links selected in the 4 clusters has distribution values similar to the LASSO model, unlike the model fitted to the links selected in the 6 clusters which have estimated values more centered on the mean when compared to the reference values. This can also be observed in the relative errors of these models. With the model fitted to the random links from the 6 cluster partitioning, the errors are more dispersed and slightly bigger than those of the LASSO model. The errors from the model based in 6 clusters are between -9% and 6% considering that 50% of the data have percentage errors between -3.5 and 1.5%. When comparing the three models presented, the best one based on the $NO_x$ estimation error is that using 4 clusters. This model has only 8 predictors and it obtained relative errors between -2.4% and 2.2% considering that half of the data have errors around -0.8% and 0.6%.

Lastly, for both pollutant emissions the partitioning method increased the goodness of fit and consequently decreased the error dispersion. All the models fitted after the partitioning were smaller than the LASSO model. As shown previously in chapter 3, the links in this network are interchangeable and do not significantly interfere in the estimation of pollutants emissions, even when considering the links selected randomly inside the clusters. The network clustering methodology is based on




(a) Density of estimated $NO_x$ values.




(b) The associated errors of daily estimated $NO_x$ emission values.

Fig. 4.11 – *Density and associated error for the estimated $NO_x$ emission values for the LASSO model and the models based on the random selection into clusters.*

the homogeneity between the links of the cluster, which means that any of the links inside can be representative of the cluster. The $CO_2$ emissions were represented by at least one link for both the clusters considered. Regarding $NO_x$, only the network divided into 6 clusters does not include the orange cluster. Because of the proportionality rules described in the beginning of the section, its size only allowed selecting a single link while the random draw finally selected a null-link. This explains why the model from this cluster has a similar but slightly higher level of errors than the LASSO model. This leads us to conclude that the representativeness of all the clusters inside the model is important for the goodness of fit and lowers the number of associated errors in the estimation.

### 4.1.2 Static/dynamic dataset

#### Lottery method: Random selection at the network scale

The second dataset studied is called static/dynamic. The LASSO method selected 77 links in this dataset to estimate the network $CO_2$ emissions for a time period of 15 minutes at the network level. The lottery method used in the static/static dataset was applied here. Figure 4.12 shows the error distribution of the $CO_2$ emissions estimation in the dynamic dataset using random draws to select links.

For all the models studied, this dataset always had outliers that greatly increased the number of errors. The outliers represented between 6 and 8% of all the observations.

Considering their error distributions, all the random selections had errors between 50% and more than -50%. To better study the error distribution, we focus on the inner part of the boxplots in order to exclude outliers. Figure 4.13 shows the same error distributions as in 4.12 without their respective outliers.

All the random selections have centered distributions with errors that can reach more than ±10% and are lower than ±20%. Also the color boxes, which represent 50% of the data, generally have errors

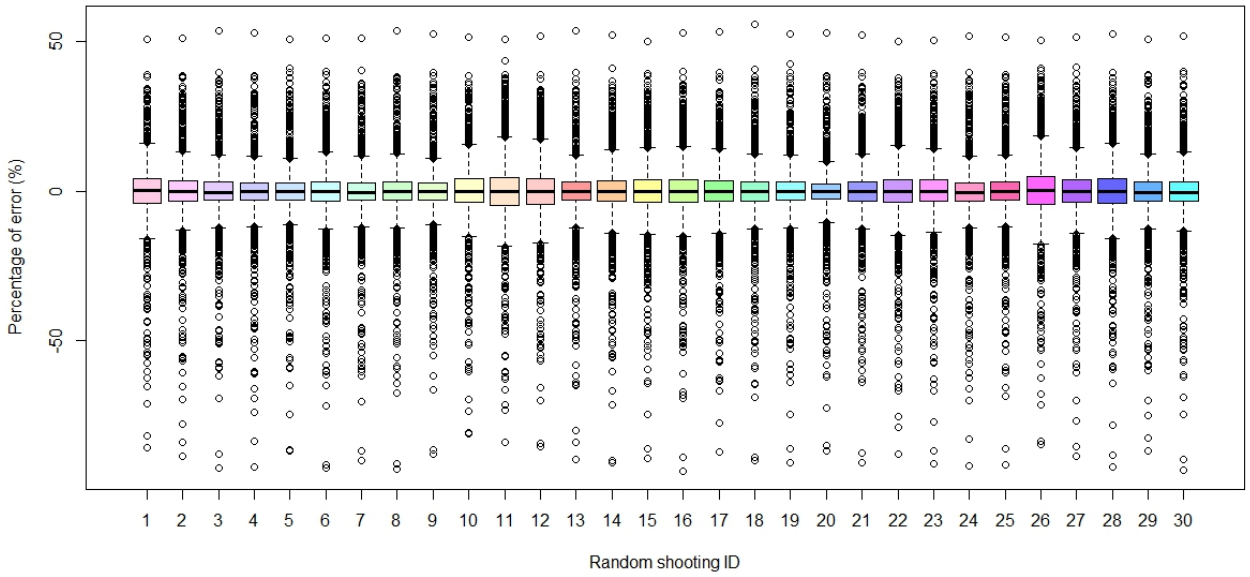



Fig. 4.12 – *The $CO_2$ error distributions for each random sampling.*

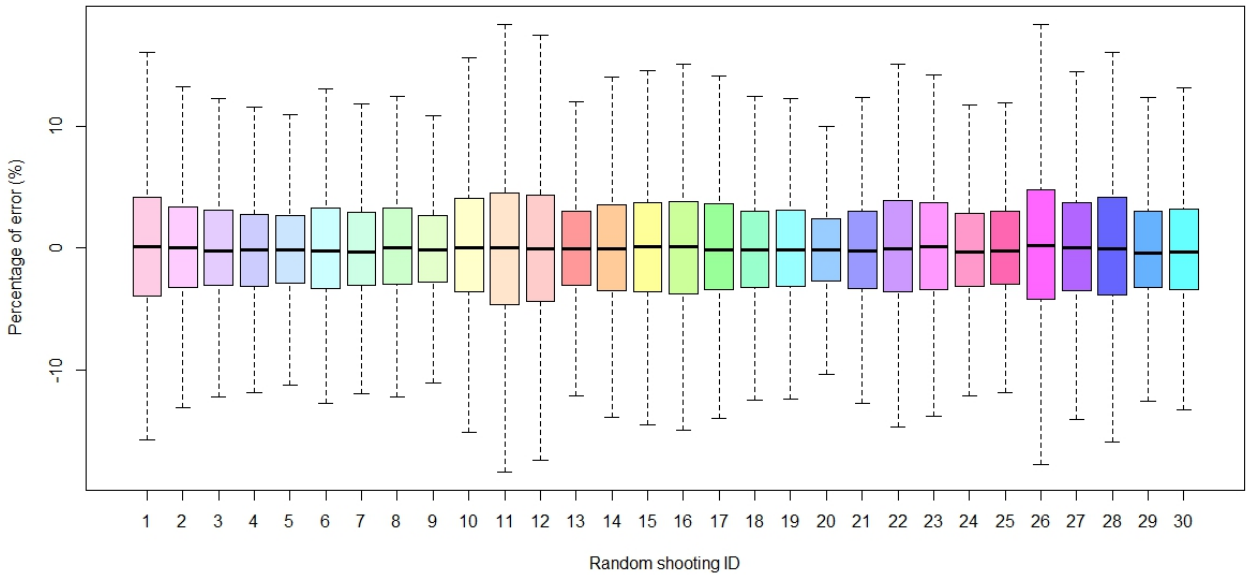


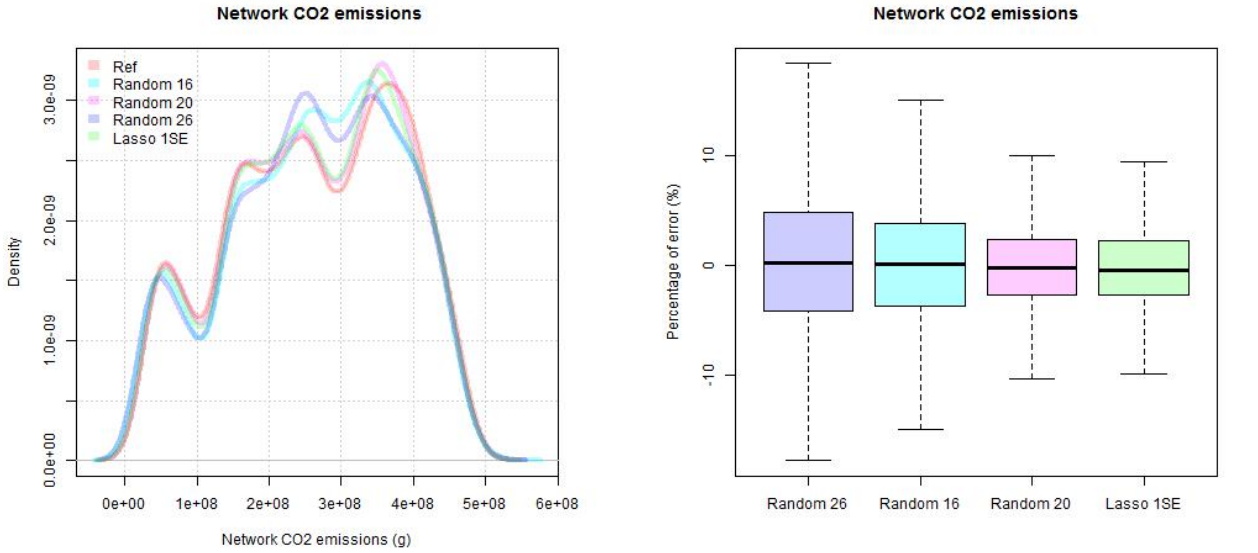

Fig. 4.13 – *The $CO_2$ error distributions without their respective outliers for each random sampling.*

between ±6%. The 1SE lambda model built by LASSO, had an error distribution from -10% to 10% and outliers that could reach over ±50% error, as shown in 3.19 in chapter 3.

Among the random draws, the best was random draw ID 20 which had fewer dispersed errors than the other ones. It also had the same level of error distribution as the 1SE lambda model. The worst case among them was random draw ID 26, with an error distribution lower than ±20%. All the draws selected between 3 and 12 null links, which means without car flows. Considering only the links with car flows, the model sizes vary between 65 and 74 links versus 77 established by LASSO. The best random draw, ID 20, has 4 null links inside the model and the worst case, ID 26, only 3. There is no linear relation between the model size and error level in this case. The majority of draws selected

around 9 null links.

The figure 4.14 shows the comparison in density of the estimated $CO_2$ emission values and the associated errors of these estimations.




(a) Density of estimated $CO_2$ values.

(b) Associated errors of estimated $CO_2$ emission values.

Fig. 4.14 – *Density and associated errors of estimated $CO_2$ emission values based on the lottery method.*

When observing the density values, the distributions of the estimated values of the proposed models are similar in comparison to the reference values. The associated errors on the estimation of these models are shown in 4.14 (*b*). All the models have between 6% and 7.3% of outliers, see the figure shown in (*b*). These outliers were omitted to ensure better analysis of the error distribution in this section. Appendix G.1 shows the same error distribution considering the outliers, but the conclusions are given in this section. Around 93% of the data are represented in (*b*) and the worst, intermediate and best models obtained using the lottery method are represented by ID 26, 16 and 20 respectively. They are compared to the LASSO model.

The model sizes from the random draws are smaller compared to LASSO. The worst has 74 non null links inside the model, the intermediate case has 71 links and the best model has 73 links. The best random draw has the same error distribution as LASSO, but with fewer predictors. Even the percentage of outliers is similar, around 7.3% of the data. The common links between the three cases of random draw and LASSO are between 33% and 38%, which proves that the models are different and not based on the same links.

The methodology used for $CO_2$ was applied to the $NO_x$ emissions. The LASSO method selected 65 links of the network to estimate the total emission in the 15 minute time period. Figure 4.15 shows the error distributions of the 30 random selections at the same rate defined by LASSO.

Figure 4.15 shows the errors distributions without considering the outliers. The error distribution with the outliers is shown in appendix G.2. The outliers represent between 7% and 8% of the data, and they can reach a percentage error of over -80% and +40%. The error distributions of 92% of the data are shown in figure 4.15. Some random draws have errors around ±15% and most of them are around ±10%. Three cases are highlighted: the worst, intermediate and the best random draws as a function of error distribution alone. They are identified by ID numbers 7, 19 and 27 respectively.

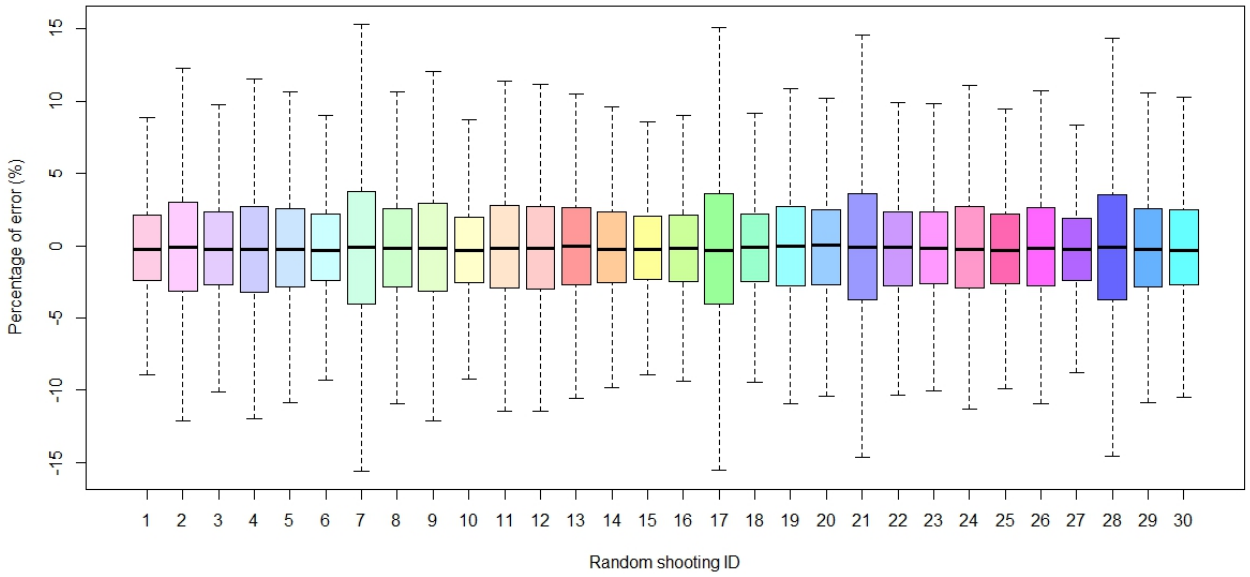
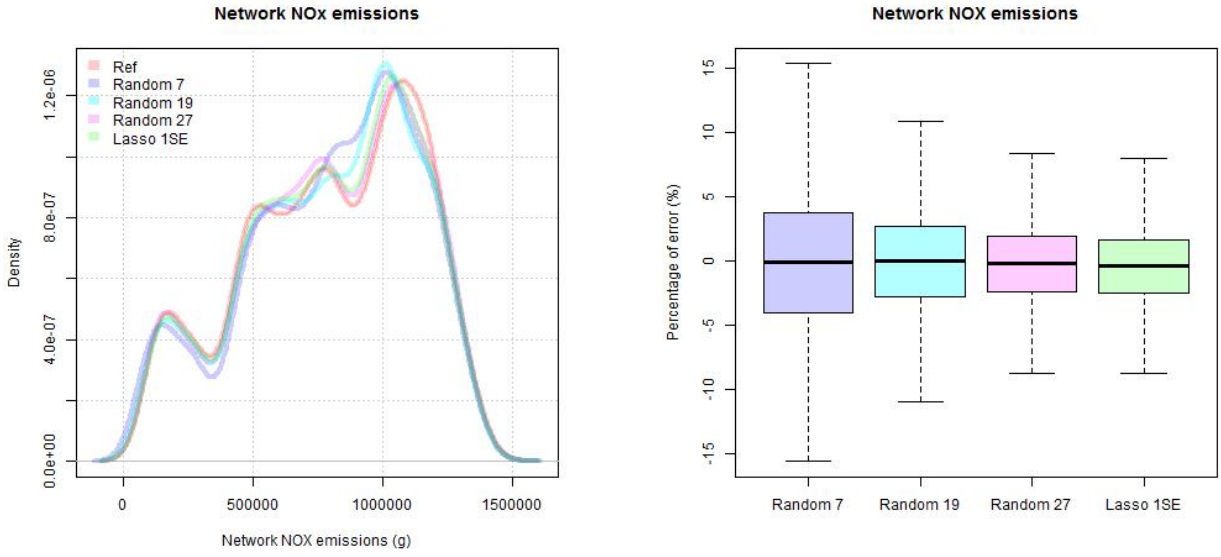



Fig. 4.15 – *The $NO_x$ error distributions without their respective outliers for each random sampling.*

All the random draws selected at least 5 null links, which means that all the model sizes were smaller than that proposed by the LASSO. The worst case was a model size equal to 56 non-null links, with the intermediate model equal to 59 and the best model equal to 60 links. The random case did not present a linear relation between the error distributions and the model sizes proposed by all draws. These three cases were compared with LASSO and the reference values. Figure 4.16 shows these comparisons and the associated errors of these models.




(a) Density of estimated $NO_x$ values.

(b) Associated errors of estimated $NO_x$ emission values.

Fig. 4.16 – *Density and associated errors from the estimated $NO_x$ emission values based on the lottery method.*

As for the $CO_2$ emissions, the densities of the estimated $NO_x$ emissions are similar to the reference values. Comparing the associated errors of these models, the best random draw has the same error

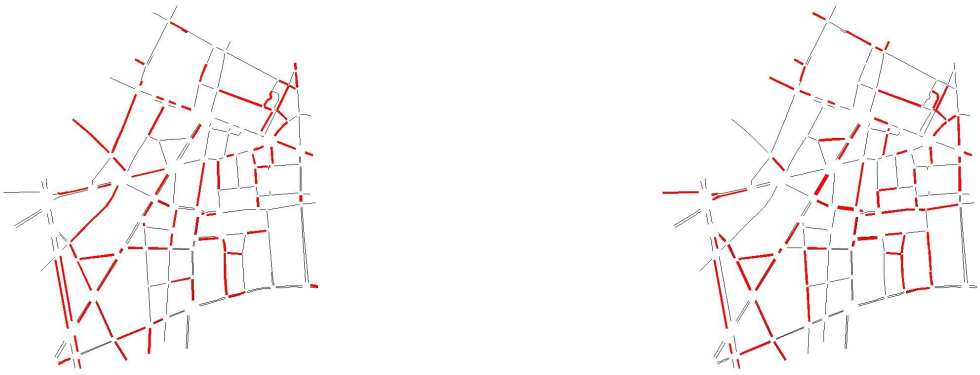
distribution and fewer predictors in the model compared to the LASSO model. The worst case can reach ±15% error. The outliers of this error distribution are presented in the appendix G.2. For these cases it represents 7% of the data on average.

For both pollutants, all the random draws have outliers and represent around 7% of the data on average. The outliers have errors over ±50% for the $CO_2$ emissions and over ±40% for the $NO_x$ emissions. Also, the variability of the errors for all the draws is between slightly less than ±10% and ±15% for both pollutants. The density values of estimated emissions present similar distribution shapes to the reference values. The error distributions of the best cases between all the draws are similar to those of the LASSO model. Also the best cases of the random draws are based on fewer links than LASSO. In general, considering the number of links needed to estimate the emissions, the random draw method is an option that can be used and it provides linear models slightly larger than LASSO.

### Random selection in a cluster

Clustering in 4 and 6 regions was maintained to observe whether the same clusters can reduce errors and increase the goodness of fit using a fine description of the traffic data.

The partitioning based on $CO_2$ emissions in 4 and 6 clusters can be seen in figure 4.6. The description and other characteristics of each cluster were explained in the previous section on the static/static dataset. The 77 links were randomly selected, based proportionally on the cluster sizes. Figure 4.17 shows the random selected links in the network divided into 4 clusters and 6 clusters.



(a) Random selected links of the network divided into 4 clusters.

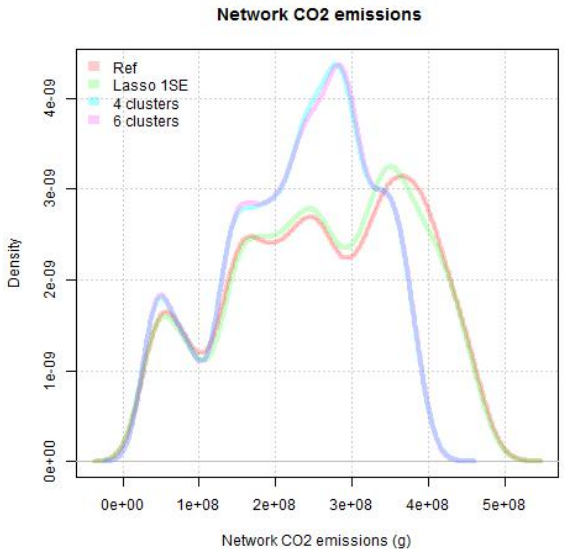
(b) Random selected links of the network divided into 6 clusters.

Fig. 4.17 – *77 links selected randomly in both cases of partitioning.*

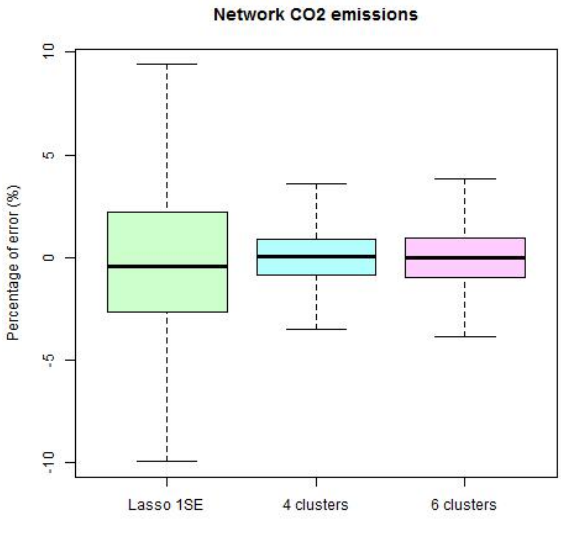
In both selections, 25 links of the 77 selected links are common, showing the difference between both partitions of the network when proportionality is taken into account. Considering the clustering in 4 zones, the random method selected 7 null-links, consequently the model is fitted to only 70 links. The second partition in 6 zones selected 8 null-links, which means a model size equal to 69 links. In both cases the fitted model is based on fewer links than defined by the LASSO method and all clusters are represented in these models.

Using the validation set, the distribution of the estimated values and the associated errors of these models are shown in figure 4.18.

Contrary to the results of the static/static dataset, the estimated values have different distributions




(a) Density of estimated $CO_2$ emission values.




(b) The associated errors of estimated values of $CO_2$ emissions.

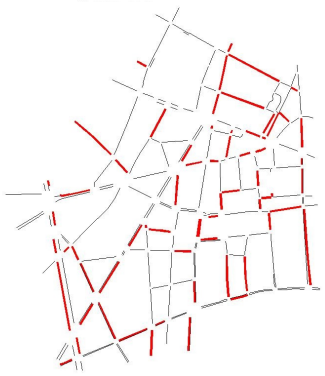
Fig. 4.18 – *Density and associated errors for the estimated $CO_2$ emission values for the LASSO model and the models based on the random selection in clusters.*

compared to the reference values. In both network partitions, the model fitted to the selected values has higher densities than the simulated values of the network used as reference. The estimated values from the model built by LASSO follows the shape of the distribution values from the reference. When analyzing the associated errors, the clustering method significantly reduces the errors in general. Considering the variability of the data, all the models have outliers and they are shown in the appendix H.1. Most of the data of the model fitted to LASSO has errors distributed between -10% and 10% and 7.3% of the data are outliers that can reach errors over -80% and over 40%. On the other hand, these values are smaller in both network partitions. The model from the partition into 6 zones has errors around ±3.8% and up to 50% of its data has errors between -1.0% and +0.9%. Also, only 2.3% of its data are outliers with errors less than 8%, as shown in figure H.1. The partitioning of the network into 4 regions provides the model with the smallest error distribution. The estimated values have errors between ±3.6% considering that 50% of its data has ±0.9% error. Also, the number of outliers is slightly lower, only 2.1% with an error lower than ±7.5%.

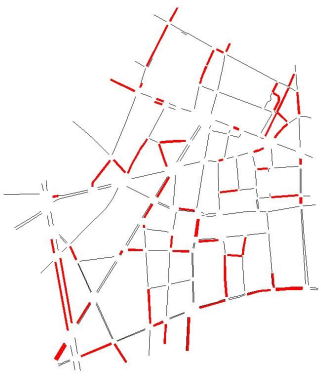
For the $NO_x$ emissions, the results are mostly similar. Figure 4.19 shows the random selected links used to estimate $NO_x$ emissions.

Only 12 of the 65 links selected in both partitions are the same. Also, fewer null-links were selected, only 4 null-links considering the partition in 4 zones and 3 null-links for the partition in 6 zones. Consequently, the models are smaller than the LASSO model and have respectively 61 and 62 links in the model. The linear model was fitted to the validation set of these links. The estimated values and the estimation error of both models are shown in figure 4.20.

Regarding the $CO_2$ emissions, the densities from the estimated values of both partitions have a different shape compared to the reference values. The estimation errors are smaller than those of LASSO. The outliers from the models can be seen in the appendix H.2. The 4-zone partition has a centered error distribution varying from ±5.6% and 50% of that of the data is between ±1.4%. For 6-zone partition the errors are even smaller. The error distribution of the estimated values is between

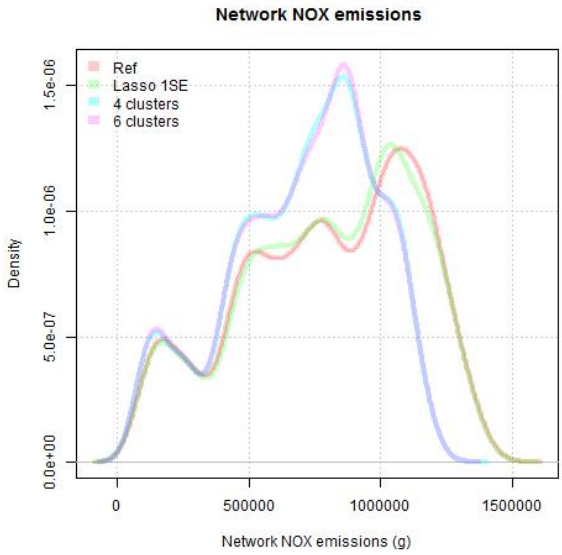


(a) The random selected links of the network divided into 4 clusters.

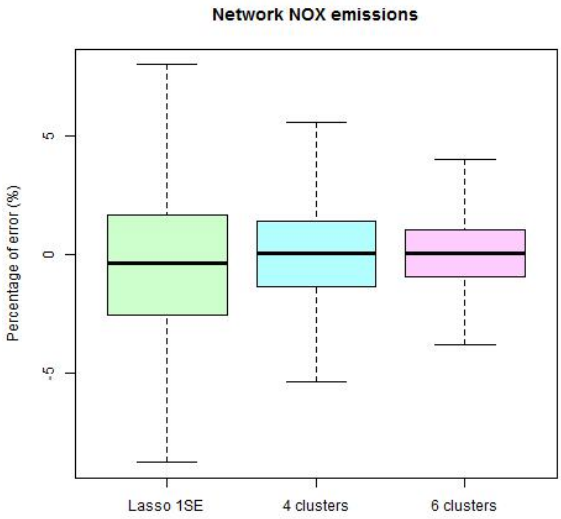


(b) The random selected links of the network divided into 6 clusters.

Fig. 4.19 – *The 65 links selected randomly in both cases of partitioning.*




(a) Density of estimated of $NO_x$ emission values.




(b) Associated errors of estimated $NO_x$ emission values.

Fig. 4.20 – *Density and associated error for the estimated $NO_x$ emission values for the LASSO model and the models based on the random selection in clusters.*

-3.8% and +4%. 50% of the data have errors around ±1.0%. The model defined by LASSO has more dispersed errors. This model presents errors between -8.8% and +8% and the outliers represent 7.9% of the data. The partitioning also reduces the number of outliers. In the 4-zone cluster the outliers represents 0.7% of its data while for the 6-zone partition they are 3.7% of the total. It is interesting to observe that the 4-zone clusters have more dispersed errors compared to the 6-zone clusters, with a few outliers.

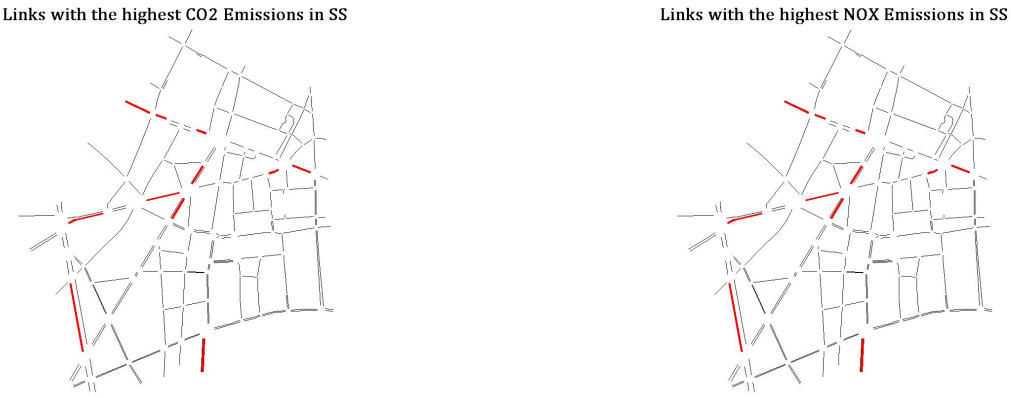
## 4.2 Ranked links

The second method, called ranked selection, is based on ranking the links of the network taking into account their values (i.e. average amount of emissions). The same number of most pollutant links will

be selected as for the LASSO 1SE lambda model for each dataset, and then a linear regression will be applied to compare the predicted values between the ranking method and LASSO. The same training and validation sets are used to fit and to validate the model. This methodology was applied for the static/static dataset and static/dynamic dataset, and for both pollutant emissions. The ranked model is different from the other methodologies applied in this chapter, but it has exactly the same number of predictors as LASSO in each dataset, as the links are ranked from the most to the least pollutant emitters, meaning no null-link can be selected.

### 4.2.1 Static/static datasets

The 11 most pollutant links are selected for each pollutant and are shown in figure 4.21.




(a) $CO_2$ emission links.

(b) $NO_x$ emission links.

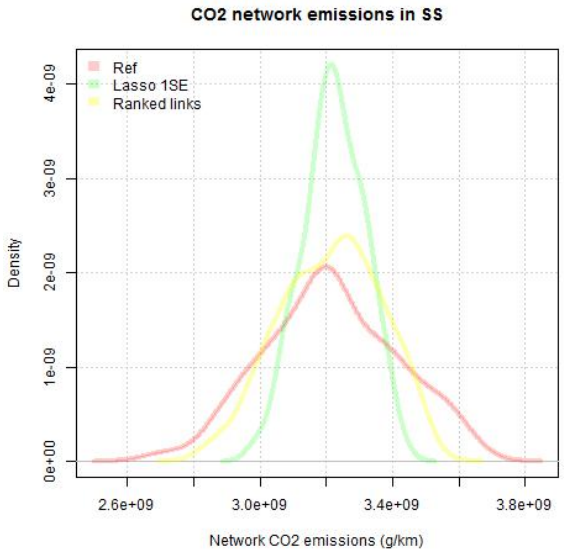
Fig. 4.21 – *Most pollutant links of the network.*

For both pollutants, the most polluted links are the same because both depend on the same traffic variables. A linear regression was applied to the emission values of these links and then the model was applied to the validation set to calculate the associated errors of these models. Figure 4.22 shows the density of the predicted values and the associated error for each model studied.

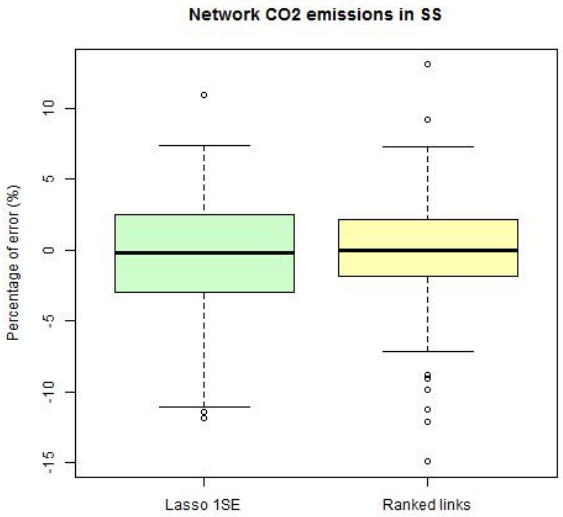
For the $CO_2$ emissions, the linear regression of the 11 most polluted links data had a density distribution similar to the reference values. When the associated errors are compared with the LASSO model and the fitted model of the ranked links, the errors from the ranked model are smaller but with a few more outliers than in LASSO.

The model built with ranked links presents the $NO_x$ emissions with almost the same density as the LASSO model. Consequently, the error distributions between both are almost similar, but the model that used ranked links has slightly narrower error dispersion than the LASSO model, with a few more outliers.

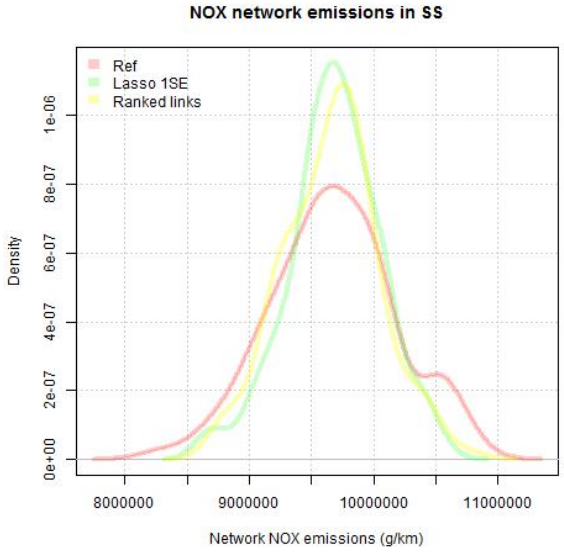
The model built using linear regression in the ranked links presents an estimation of daily emissions for both pollutant emissions, and they had fewer errors compared to LASSO. It is interesting to observe that using only one set of links, it is possible to estimate both pollutant emissions with the same selection rate and the same level of estimation error as LASSO. For the models studied previously, each pollutant has its own set of links with some in common when both pollutants are compared. For this district of Paris, analyzing the daily emission values and ranking them allows estimating the pollutant emissions using a single set of links. More than 95% of the predicted values have percentage



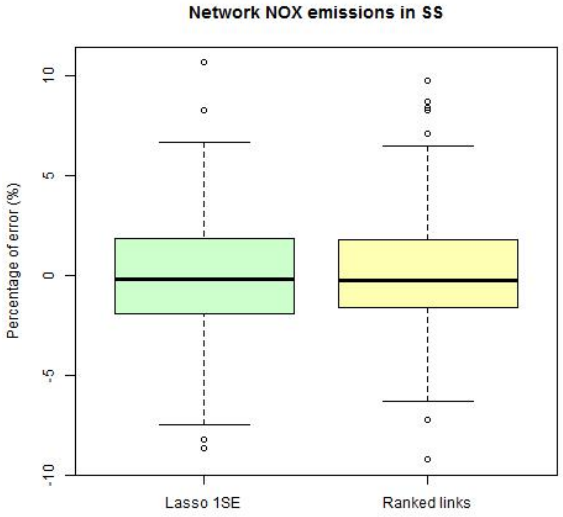

(a) $CO_2$ emission densities.




(b) $CO_2$ emission error distributions.




(c) $NO_x$ emission densities.




(d) $NO_x$ emission error distributions.

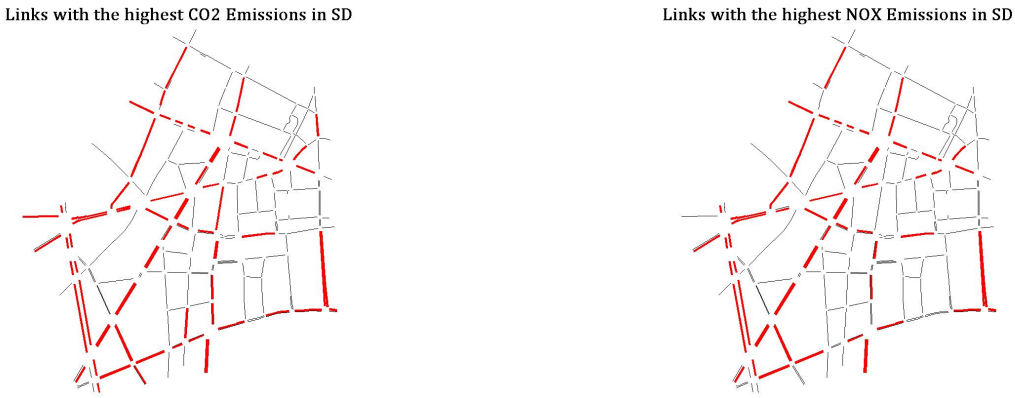
Fig. 4.22 – *Comparison between predicted values.*

errors lower than -6.5% and 6.5%. Also, 50% of the data have percentage errors lower than ±2%. The outliers in LASSO represent 3% of the data and only 5% in the ranked model.

By identifying the most pollutant links on the network, and using less than 5% of the links, daily emissions can be estimated for both pollutants with the same level of error in comparison to LASSO. Only a few situations cannot be estimated with a confidence interval of 95% .

### 4.2.2 Static/dynamic datasets

Because of the high variability of the data at link level used to estimate the network emissions on this temporal scale, the 1SE lambda built by LASSO selected 77 links of the network to estimate $CO_2$ emissions and 65 links to estimate $NO_x$. As with static/static dataset, the same amount of links in each pollutant is selected after ranking from the most to the least polluted links, see figure 4.23.

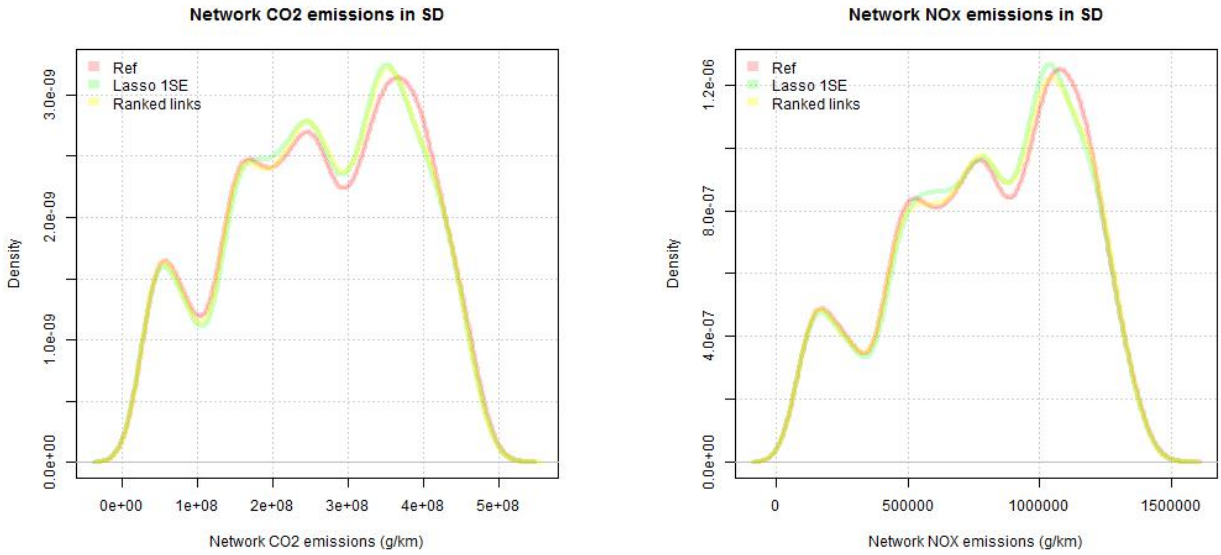



(a) $CO_2$ emissions links.

(b) $NO_x$ emissions.

Fig. 4.23 – *Most pollutant links of the network.*

When the links were compared between both pollutants, all the links selected for $NO_x$ were within those selected for $CO_2$. A linear regression was applied to each pollutant and the predicted values were compared with those from the optimal model of LASSO. The densities of the predicted values from both models and the reference ones are shown in figure 4.24.



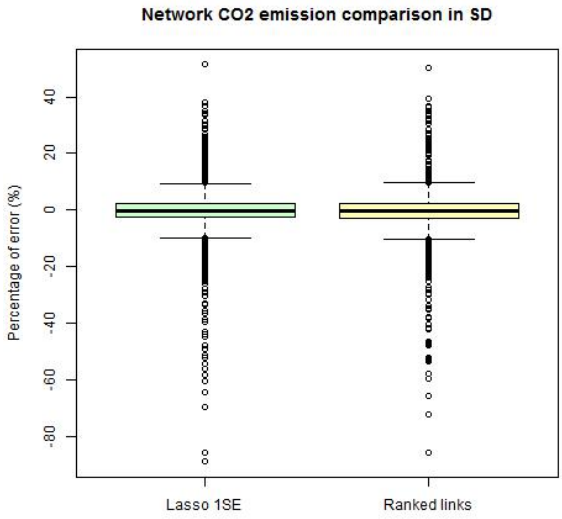

(a) $CO_2$ emission densities.

(b) $NO_x$ emission densities.

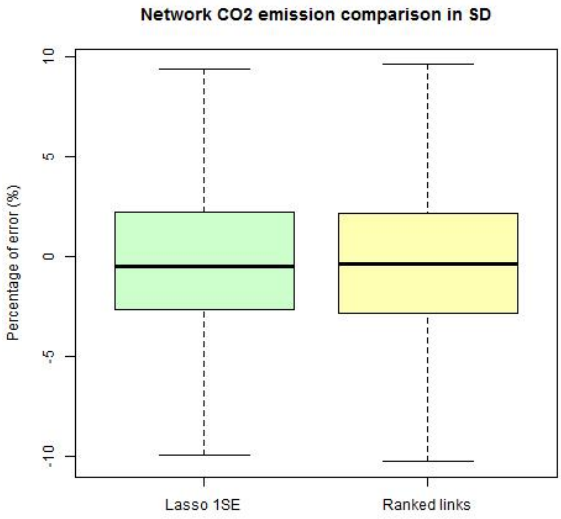
Fig. 4.24 – *Densities of emission values.*

Both models are quite close compared to the reference values. The relative error values are calculated to analyze the associated error of each model. The error distributions are shown in figure 4.25.

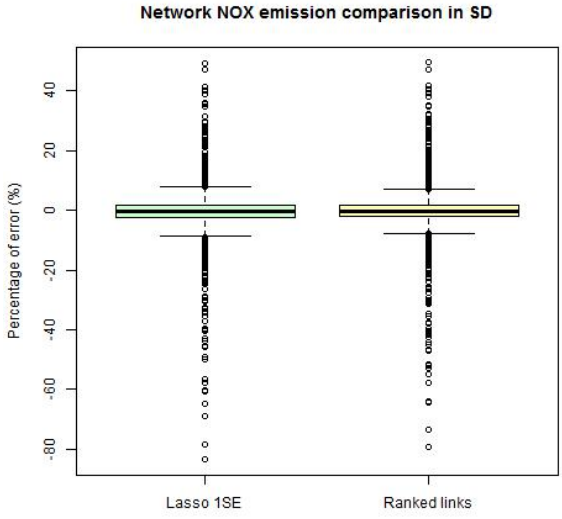
Comparing both pollutant emissions it can be seen that in $(a)$ and $(c)$, the errors can reach -80% and +40% and their distributions are similar. For the $CO_2$ emissions, the outliers from the 1SE lambda model represent 7.3% of the data and 7.1% in the ranked model. The $NO_x$ emissions present a few more outliers in both predictive models; the 1SE lambda model has 7.9% outliers compared to the ranked links with 8.2% outliers. More than 90% of the data for $CO_2$ emissions have percentage errors



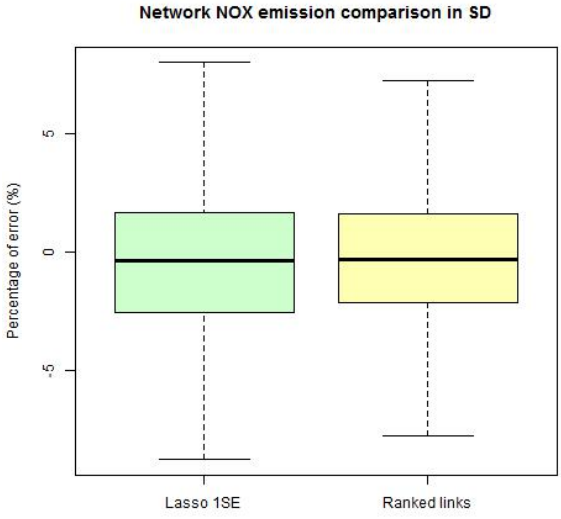

(a) $CO_2$ emission error distribution.




(b) $CO_2$ emission error distribution without outliers.




(c) $NO_x$ emission error distribution.




(d) $NO_x$ emission error distribution without outliers.

Fig. 4.25 – *Error distributions from the predicted values.*

lower than ±10% for both models, and 50% of the data has a percentage error between -2.8% and +2.2% on average for both models. The model based on the ranked links has slightly more dispersed errors; around 0.3% more than 1SE lambda. The $NO_x$ has a lower number of errors compared to $CO_2$. The 1 SE lambda model has between -8.8% and +8% errors and the model based on ranked links has between -7.8% and +7.2%. More than 90% of the data are within this error range considering that 50% of them have a percentage error of -2.5% and less than +2% on average. The model built with ranked links presents less dispersed errors than the 1 SE lambda model, with an average difference between 0.8 and 1%.

In general, with the same rate selection using the most pollutants links in the network, a linear model can be built estimating emissions with the same level of error compared to the LASSO. From these models, less than 10% of the data cannot be estimated considering a confidence bounds of 95%;

they are represented by the outliers. Using daily values and identifying the most pollutants links of the network, it is possible to estimate the emissions with reasonable error needing only data for 5% of the network links. Both models present good predictions for the district studied when the time range is reduced to estimate emissions for each 15 minutes.

## 4.3 Stepwise selection

Stepwise regression is a linear regression method like the previous proposed methods, but the selection of the predictive variables is based on an automatic strategy (Efroymson, 1960) and (Flomn and Cassell, 2007). The method starts by adding and removing terms from the linear model automatically, based on their statistical significance in the linear regression. The method builds an initial model to start the procedure, and then compares the accuracy to the overall prediction of larger and smaller models. At each step when a variable is added or removed from the model, the $p$-value of an $F-$statistic (i.e. the probability of F) is calculated to test the model with and without the variable. Each variable is tested by the null hypothesis. This means that if the null hypothesis is accepted the variable is added to the model, if not it is rejected (Draper and Smith, 1998). Conversely, if a variable is included in the model, the null hypothesis means that the variable has a coefficient equal to zero. If the evidence cannot reject the null hypothesis, the variable is removed from the model.

The first algorithm was proposed by (Efroymson, 1960). Basically the methodology uses the probability of F, which means the p-value, or the F value to include a variable in, or remove it from, the equation. The criteria that controls the inclusion or removal of a variable is determined by the probabilities of "$F$-to-enter" and "$F$-to-remove". The default values of the significance level used are 0.05 “to enter” and 0.10 “to remove”. After defining the thresholds the statistical selection procedure can start by adding a variable which has the smallest $p$-value among all the variables available. Then it continues by entering another variable that is not inside the model with a smaller $p$-value for $F$ and so on. No matter which variable is added to the model, it is impossible to know which one will be included or removed from it(Hastie et al., 2013).

The stepwise procedure plays a key role in performing model selection through $F-$statistic $= t-$squared. The three main steps of the stepwise procedure according to (Department of Statistics Online Programs, The Pennsylvania State University, 2016) are described below. Let $y$ be the response value that must be estimated by the model, and $x_n$ be the variable identification from 1 to $n$ available variables. The steps are described as follows:

- Step 1: Fit an initial model to each variable, so each model contains only one predictor. Mathematically speaking, this means that regress $y$ (the response values) is fitted to $x_1$ (the variable), then regress $y$ is fitted to $x_2$ and so on until the last variable. The $t$-test $p$-values are computed for all the variables. Those which have lower $p$-values than the criteria are eligible for entry in the stepwise model. Then, the first variable added to the stepwise model is the predictor that has the smallest $p$-value of all the candidates. The procedure stops if no variable has a $p$-value smaller than the established criteria.

- Step 2: It assumed that $x_1$ has the smallest $p$-values and that it is lower than the threshold fixed, consequently it is considered the best variable for inclusion in the stepwise model for the first step. Then a new fit starts considering two variables, which includes $x_1$ and the other variables that are not in the model. In other words, a new regression starts, regress $y$ is fitted

to $x_1$ and $x_2$, regress $y$ is fitted to $x_1$ and $x_3$ and so on. The second variable that will be added to the model is that with the smallest $p$-value and it must be lower than the threshold fixed to enter. If no variable has a $p$-value below the significance level to enter the model, the procedure stops. Thus, the stepwise model is obtained from the first step of the final model considered. However, this supposes that variable $x_2$ has the smallest $p$-value and is lower than the model inclusion criterion. Consequently this variable is added to the stepwise model. The procedure then reverses and checks how the $x_2$ variable can affect the significance of the first predictor, in this case $x_1$, after including $x_2$. Thus, the procedure checks the $t$-test and $p$-values using the null hypothesis. If the $p$-value is greater than the limit established, the evidence cannot be rejected by the null hypothesis, and variable $x_1$ is removed from the stepwise model;

- Step 3: It is presumed that $x_1$ and $x_2$ are in the stepwise model. The procedure starts again by fitting the $x_1$ and $x_2$ as predictors with the other variables not yet in the model. That is to say, it steps back to y on $x_1$, $x_2$ and $x_3$, then steps back to $y$ on $x_1$, $x_2$ and $x_4$, and so on until stepping back to $y$ on $x_1$, $x_2$ and $x_n$. As with the previous steps, all the predictors with $p$-values lower than the criteria are possible candidates for entry in the stepwise model. The predictor with the smallest $p$-value is included in the model. If any variable has a $p$-value lower than the criteria, the procedure stops and the final model obtained will include only two variables. But, if we consider that variable $x_3$ meets all the criteria for inclusion in the stepwise model, the procedure steps back and checks if the inclusion of variable $x_3$ in the stepwise model affects the $p$-values of $x_1$ and $x_2$. The null hypothesis evaluates if $x_1$ or $x_2$ or both become irrelevant; if so one or both are removed from the stepwise model, and then the procedure starts again. The procedure continues these steps until any combination of variables can be in or out, which means any combination of variables can have $p$-values below or above the entry threshold and the process is stopped when the final model has been obtained.

For each step the $t$-statistic is calculated for the variables inside and outside the stepwise model. For the variables inside the model, the t-statistics are calculated for the estimated coefficients, then they are squared and used as $F$-to-remove. The $t$-statistic is also calculated for the variables outside the model and used as the coefficient of the variable if this variable is the next one to be added in the stepwise model, then squared, and considered as $F-$to-enter (Nau, 2017). The combination of the $t$-statistic, the $p$-values and the threshold to enter or leave the model forms a procedure based on the probability of variable selection.

The stepwise process can be biased depending on the data and the criteria fitted, because it focuses only on one step at a time, forward or backward, at any point (Derksen and Keselman, 1992). The stepwise method can build different models from the same group of variables depending on the variables included in the stepwise model when the procedure starts. The stepwise method stops when no step can improve the model.

Taking into account the considerations above, some aspects of this method should be recalled when using it. The final stepwise model is not necessarily the optimal model with the largest number of variables inside (Mark and Goldberg, 2001). The procedure can give many equally good models using the same set of variables as it depends on which order the variables are included or excluded. Also, there is no guarantee that the process includes only the important variables and excludes the unimportant ones. As concluded by (Department of Statistics Online Programs, The Pennsylvania State University, 2016) through examples, many $t$-tests were conducted for testing the null-hypothesis

during the steps, hence there is a high probability that certain non-relevant variables could be included in the model while other important ones are excluded.

As with the other methods applied here, the stepwise method was applied to the two main datasets, namely static/static and static/dynamic, and the observations were divided into a training and a validation set for each one.

The same partition of the data set built in the LASSO study was used to apply the stepwise regression. This means that for each data set, the same set as that used to fit a model to LASSO (2/3 of the dataset) and the set used to validate the fitted model (1/3 of the dataset) will be the same for the stepwise selection. The method selects the links of the network and provides a model to estimate the network emissions on two temporal scales. Then, the associated errors are calculated. Both methods, LASSO and stepwise, are compared. Only the pollutant emissions from static/static dataset (daily values) and static/dynamic dataset (network 15 minutes values) are studied.

### 4.3.1 Static/static datasets

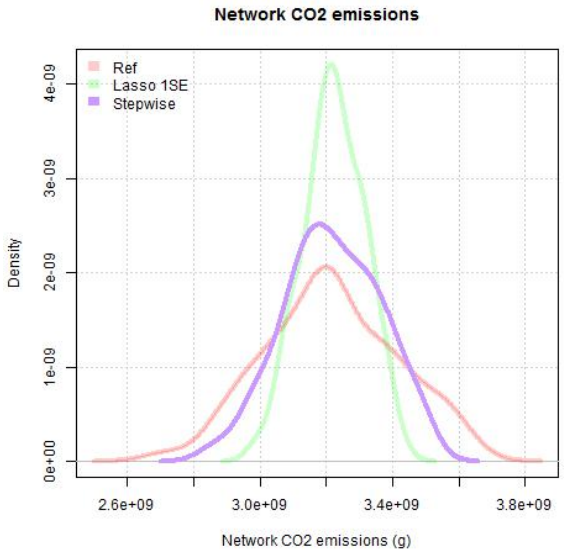
11 links were selected for the $CO_2$ and $NO_x$ emissions using the optimal model from LASSO. The same set used for training and validation in LASSO was used to fit and validate the stepwise selection model to facilitate the comparison between them. Figure 4.26 shows the density of predicted values from both models and their associated errors.

For both pollutant emissions, the values predicted by the stepwise method are closer to the reference values than the ones predicted by LASSO, mainly in $CO_2$. Also for both emissions, the stepwise method provides a smaller model than that defined by LASSO and with less dispersed associated errors.

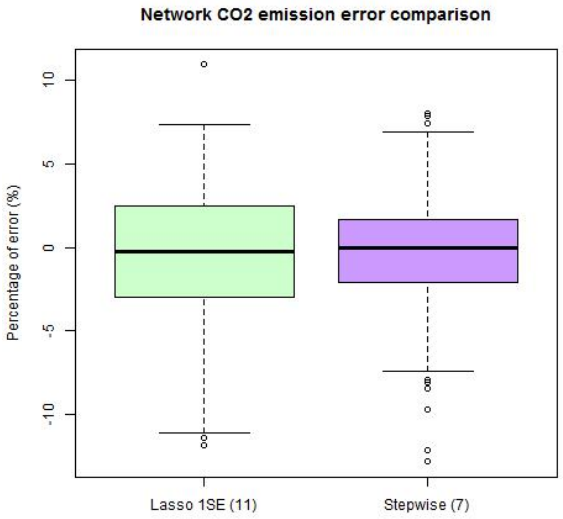
For the $CO_2$ emissions, the stepwise method selected 7 links in the network, which corresponds to 3% of the total number of links. More than 90% of the data had errors between -7.4% and 6.9%, and 50% of the data had errors between -2.1% and 1.7% with a median value of -0.02%. The LASSO model presented more dispersed errors based on 11 selected links (5% of the network). More than 95% of the data in the LASSO model had errors between -11.1% and 7.4%, and 50% of this 95% had errors between -3% and 2.5% with a median value of -0.2%. The LASSO model presented a right-skewed distribution and the stepwise a more centered one. Both models presented few outliers. In the LASSO model they represent 2.3% with extreme values with a percentage error of around 11%. The stepwise model has more outliers than LASSO, which represents 6.8% of the data with extreme values that can reach -13% and 8% error.

Considering the pollutant $NO_x$, the stepwise method selected 8 links (3.5% of the network), 3 links fewer than LASSO. More than 90% of the predicted values in the stepwise model had associated errors of $\pm$5.3% and 50% of the data inside it had $\pm$1.3% with a median value of 0.07%. The predicted values of the LASSO model had associated errors in comparison to the reference values of between -7.5% and 6.7% and referred to more than 95% of the data. Of this this 95%, 50% had associated errors between $\pm$1.9% with a median value of -0.2%. The stepwise model provided a few more outliers than LASSO. The outliers from the stepwise model were represented by 8.3% of the data versus LASSO with only 3%. In both cases, the extreme values of the outliers were around $\pm$10%.

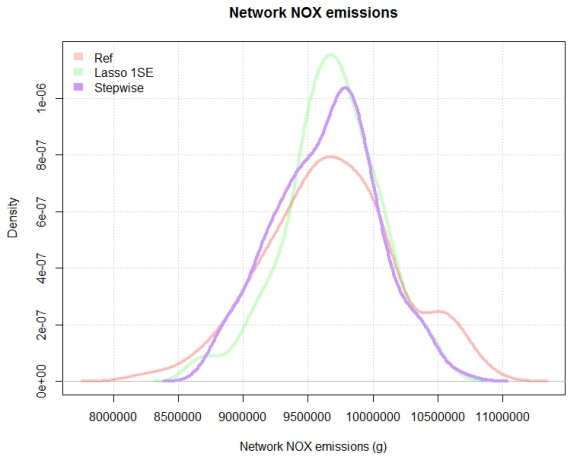
Considering the selected links of both proposed models, 72.7% of the $CO_2$ and $NO_x$ links were common in the LASSO selections. For the stepwise selection, both pollutants had only 2 selected links in common, which represents less than 29% of the $CO_2$ model and 25% of the $NO_x$ model. This shows that each pollutant has its own specific model of estimation with different model sizes defined by the



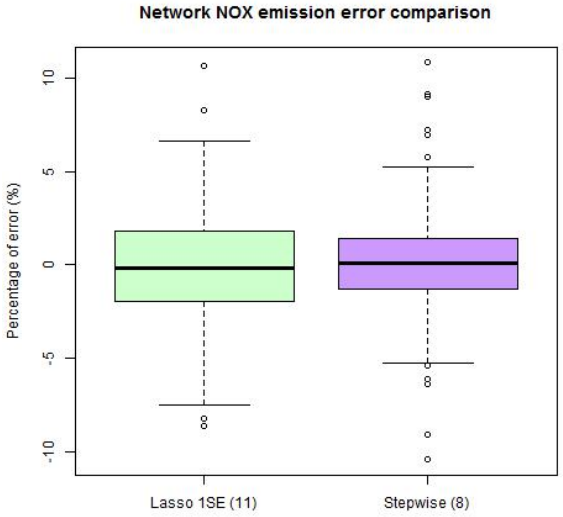

(a) $CO_2$ emission densities.




(b) Error distributions of$CO_2$ emissions.




(c) $NO_x$ emission densities.



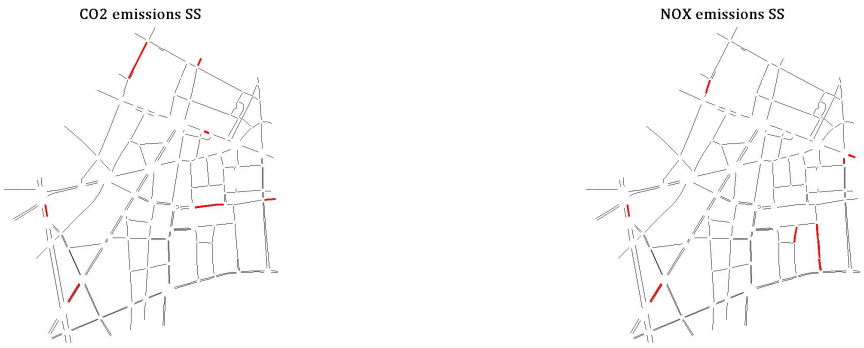

(d) Error distributions of $NO_x$ emissions.

Fig. 4.26 – *Densities and error distributions from the predicted values by each model.*

stepwise procedure. If the $CO_2$ links are compared between the LASSO and stepwise selections, they have only one link in common, which means that the stepwise model is completely different from the LASSO model with a smaller set of links and with fewer associated errors. The same occurs for $NO_x$, as the LASSO and stepwise selections have only 3 links in common, which represents 37.5% of the stepwise model which is smaller and has less error dispersion than the LASSO model. Figure 4.27 shows the links selected by the stepwise method for each pollutant.

### 4.3.2 Static/dynamic datasets

The optimal model from the LASSO method selects 77 links to estimate the network $CO_2$ emissions for a time period of 15 minutes and 65 links for the $NO_x$ emissions. For all the analyses, the figures that show the outlier distribution for all the models were placed in the appendix while the conclusion




(a) $CO_2$ selected links.

(b) $NO_x$ selected links.

Fig. 4.27 – *Links selected by the stepwise method in the static/static dataset.*

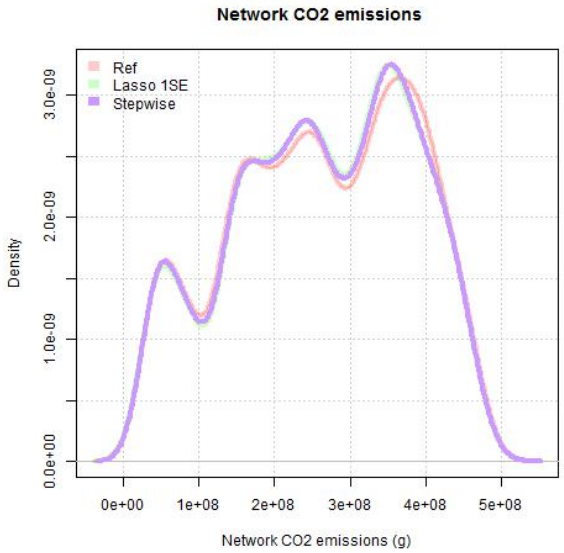
is provided in this section. Figure 4.28 shows the densities and error distributions of the LASSO and stepwise models for comparison considering both pollutant emissions. Figure I.1 shows the same figure but considering the error distribution of outliers.

The values predicted by both models present similar densities to the reference values for both pollutants. For $CO_2$, the associated errors from the LASSO and stepwise models are very similar, but their model sizes are different. The stepwise model presents an error dispersion rather less scattered than that of LASSO. More than 92% of the data has errors between -9.6% and 9%, considering that 50% of this data lie between -2.5% and 2.1% with a median value of -0.3%. Similarly, more than 92% of the associated errors of the values predicted by the LASSO model are between -10% and 9.4%, considering that 50% of them are between -2.7% and 2.2% with a median value of -0.5%. Both models have outliers as associated errors and they represent 7.3% of the data in each model. Their extreme values are similar in both models, with percentage errors of around -89% and 51%.

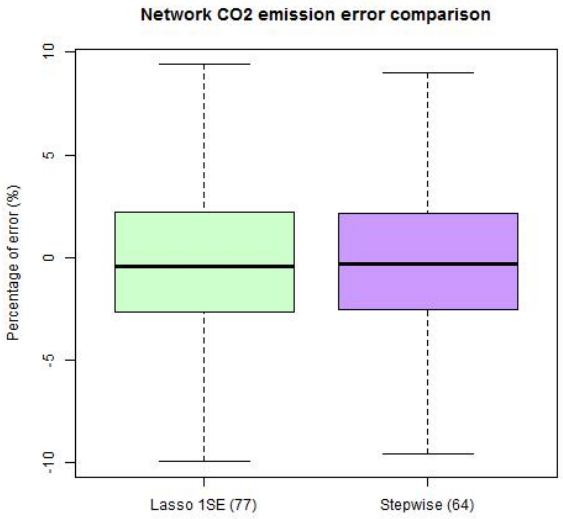
The associated errors of the stepwise model in $NO_x$ are more visible and also less dispersed in comparison to the LASSO model. Considering that 91% of the data from validation set have errors between -7.2% and 6.8% taking into account that 50% of this data -2% and 1.5% with a median value of -0.3%. In contrast the LASSO has 92% of data between -8.9% and 8.0%, 50% from it are between -2.6% and 1.7% with median value -0.41%. Some predicted values from both models are considered as outliers and they are shown in I.1. They represent 7.8% of the validated set in the LASSO model and 9% in the stepwise model. For the first model the outliers can reach -83.4% and 49.2%, and in the stepwise they reach -80.0% and 50.0%.

The size of the stepwise model is smaller for both pollutants than the LASSO selection. The links selected by the stepwise model for $CO_2$ and $NO_x$ emissions are shown in figure 4.29.

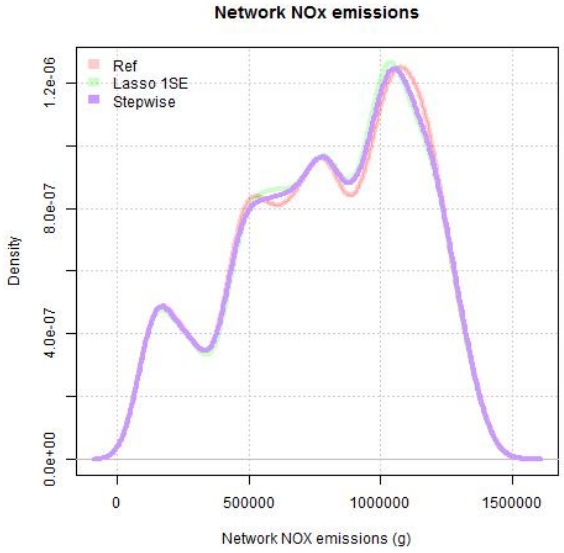
The stepwise method has a selection rate of 27.8% for $CO_2$ and 26.5% for $NO_x$. They have in common 45 links, which represents 70.3% of the $CO_2$ selection and 73.8% of $NO_x$ model. When the LASSO and stepwise link selections are compared for each pollutant, they have 44 links in common for $CO_2$ but only 31 links in common for $NO_x$. The stepwise model selected at least 30% network links different from those selected by LASSO, which provided a better estimation of the network emissions for a 15 minute time period considering fewer dispersed associated errors.



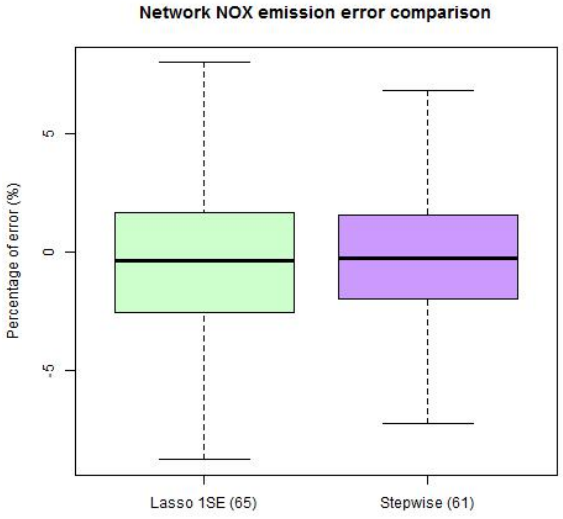

(a) $CO_2$ emission densities.




(b) Error distributions of$CO_2$ emissions.




(c) $NO_x$ emission densities.



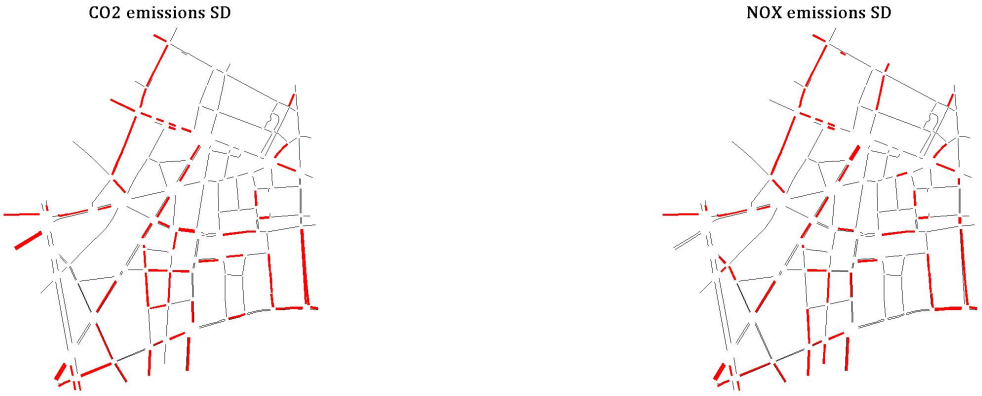

(d) Error distributions of $NO_x$ emissions.

Fig. 4.28 – *Densities and error distributions from the predicted values by each model.*

### 4.3.3 Conclusion

The aim of this study was to compare two different smart selection methods based on a linear regression to select the most relevant links of the network. The first was the LASSO method which was studied in chapter 3, and the second method, called the stepwise method, was presented in this section. Two temporal ranges were studied: (i) daily emission values of the network and (ii) network emissions for time periods of 15 minutes. Also, two pollutants were considered, $CO_2$ and $NO_x$.

Comparing both sampling methods, the stepwise method selected fewer links of the network in all cases. Also, the associated errors were less dispersed than those of LASSO. The static/static dataset represented the daily values of the network. Although the stepwise method had a lower sampling rate than LASSO, it needed at least 27% of the network to build a model with a reasonable estimation error for the static/dynamic dataset. Only 50% of the links selected for $CO_2$ and $NO_x$ were in common.




(a) $CO_2$ selected links.

(b) $NO_x$ selected links.

Fig. 4.29 – *Links selected by the stepwise method in the static/dynamic dataset.*

This study clearly showed that it is possible to estimate emissions at network scale using a set of links, i.e. a few links. The stepwise method presents a better solution in terms of estimation error and sampling rate.

## 4.4 Comparison of all the methods

In the previous sections, several methods were proposed to estimate network emissions over two temporal ranges, daily and 15 minutes. All these methods were compared to the LASSO method detailed in chapter 3. The results from all these methods are compared in this section.

The first method presented is random selection. Two methodologies were proposed: the lottery method considering the entire network and a lottery method inside clusters. In both cases the number of selected links were the same as proposed by LASSO for each dataset. Chapter 3 concluded that a large number of link combinations exist to estimate network emissions without affecting the estimation. The values estimated for both the datasets and the pollutants had the same level of error as LASSO. The mean absolute error for the 30 random draw selections were compared to observe the variation of the mean error value. Figure 4.30 shows the mean absolute error for $CO_2$ emissions for each random sampling of each dataset.

The static/static dataset presents a more stable average error for all the random selections than the static/dynamic one. For all of them, the mean error was around 3%. Considering the static/dynamic dataset, the errors were more scattered, between 4% and 6%, and bigger than for the static/static case.

The comparison of datasets for the $NO_x$ emissions presented the same characteristics and considerations as for the $CO_2$ emissions. The only difference was the average error which was a little smaller in each dataset. Figure 4.31 shows the average absolute error of $NO_x$ emissions in both datasets.

Reducing the temporal scale of the data tends to increase the number of errors and can lead to more situations that cannot be described by the model (outliers). Also, the number of links selected increases with temporal reduction. All the models presented had at least one null-link selected, which means that all the models considered fewer links than those selected by LASSO. LASSO automatically excluded all null-links when building the models, contrary to the lottery method.

The second method was the random selection inside clusters. This method was based on the

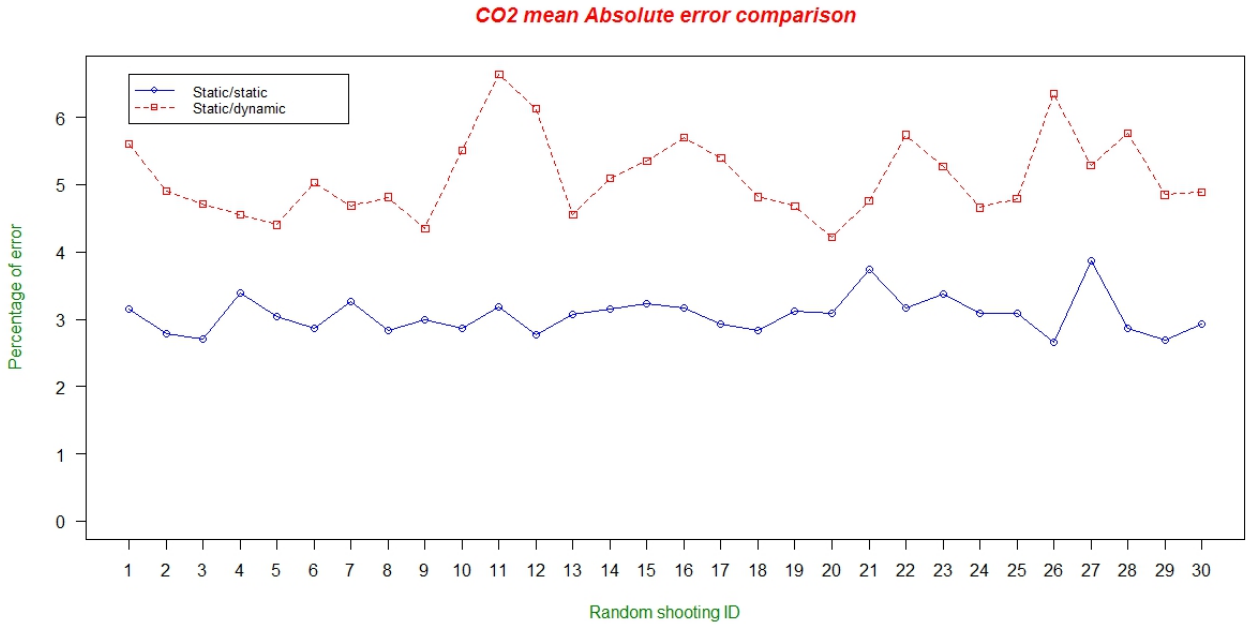



Fig. 4.30 – *The mean absolute error for each random sampling and dataset.*

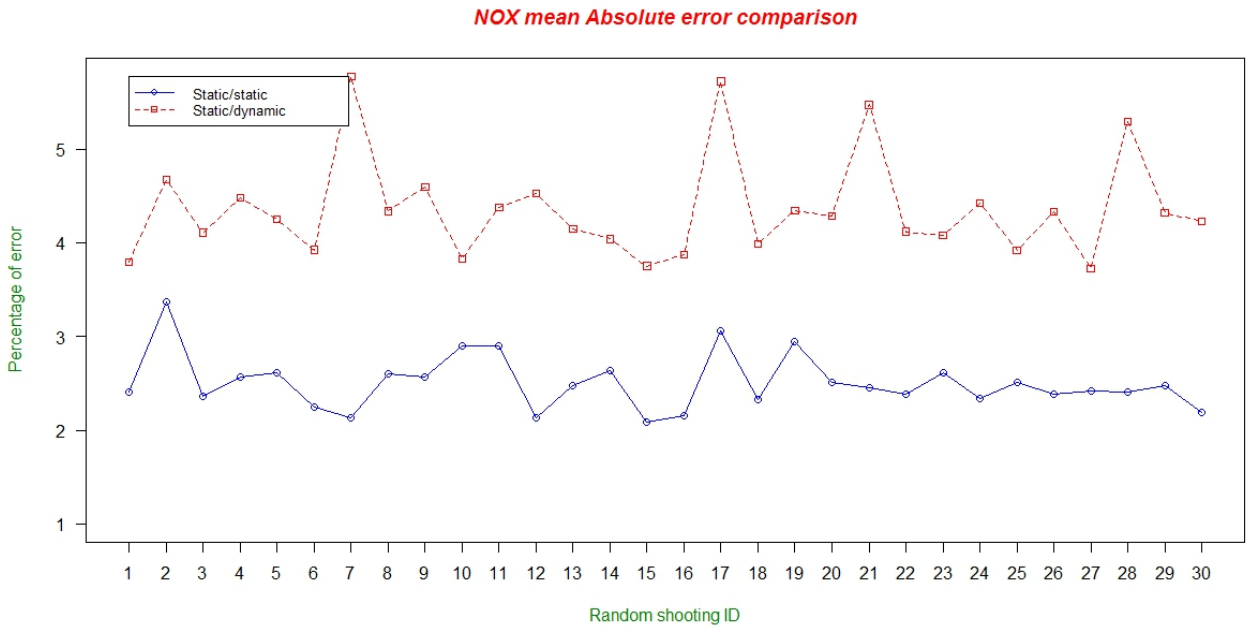



Fig. 4.31 – *The mean absolute error for each random sampling and dataset.*

partitioning of the network into smaller networks which were then analyzed iteratively to define the clusters based on their emission values. The aim was to reduce the network into some clusters based on the emissions values, to get together all links that have similar emissions behavior into the same cluster. After the clusters were defined, a random selection was performed at the same rate as LASSO as a function of the size of the cluster. This approach reduces emission estimation errors on selected links which belong to a cluster and which can be more representative of the entire group because their similarity. This method allows selecting links to represent clusters and estimate the network emissions.

In general, both pollutants have the same cluster partitions in each dataset, which is positive because it means both pollutants can be estimated with the same partition. Also, in all cases clustering

significantly reduces the error distribution compared to LASSO. Also, all the models are smaller than LASSO because of the null-links that can be selected by the random method. This trend is confirmed if all the clusters are represented in the final model. Regarding the case of $NO_x$ in the static/static dataset, the model gives bigger errors for the 6-cluster network than LASSO because the links selected in the clusters were null-link.

The third method presented was the ranked links. The main idea is to rank the emission values of network links from to the most to least pollutant. The 11 selected links represent 21% of the emissions of the network for daily traffic on average. The errors of the model in the ranked links are less dispersed than LASSO. This dataset has almost 7 times fewer predictors inside the model than the static/dynamic dataset. When the temporal range is reduced, the variability on the data is incorporated and consequently the model requires more predictors to obtain a good estimation of the emissions. The size of the static/dynamic dataset model is equal to 77 predictors for $CO_2$ and 65 for $NO_x$. The model based on ranked links has less dispersed errors than the optimal model of LASSO. The ranked links for $CO_2$ represents 76% of the total emissions from the network and the ranked links of $NO_x$ represent, on average, 70% of the network emissions for a time period of 15 minutes.

The fourth and last method is the stepwise method. It is a smart method like LASSO. For all cases the stepwise method selected fewer links and had the errors were less dispersed than for the LASSO method.

To measure the efficiency of the models to choose the best among them, the Bayesian Information Criterion ($BIC$) (Schwarz, 1978) was used to select the model that best fits the data. The $BIC$ score was calculated for each model built for the $CO_2$ and $NO_x$ emissions in both datasets. Table 4.1 shows the $BIC$ score for all the models studied.

The $BIC$ scores are similar for the $CO_2$ models using the static/static dataset. The $BIC$ methodology states that the model with the lowest score is the best fit for the data, and in this case this model is the stepwise selection. The Random ID 3 has the second lowest $BIC$ score, followed by the LASSO 1SE model, and the "6-cluster" and the ranked links models. Considering the $NO_x$ emissions in the same dataset, the ranked link model has the lowest score followed by the Random ID 15 models ( the best model of all the random draws) and then by the stepwise model.

Considering the static/dynamic dataset, for both pollutant emissions, the stepwise model has the lowest $BIC$ score, followed by the LASSO 1SE and ranked link models. It is interesting to observe that for all the models proposed for both pollutants and both datasets, the stepwise method almost always has the best $BIC$ score or has among the three best scores compared to the other models. The other two models with similar scores for this dataset are the LASSO 1 SE and ranked link models. As can be observed, the best $BIC$ score is obtained by two smart sampling methods and by a naive method. The estimated error distributions are quite similar between these three models, even with different model sizes. The naive method considers only the most pollutant links on the network with the same selection rate as LASSO, 77 links for $CO_2$ and 65 links for $NO_x$. This method can estimate network values in any dynamic range of the network for a low computation cost and it is easy to use. But the stepwise model requires fewer links in the network for the same computational cost and it obtains an estimation with a less dispersed associated error than the other two.

In general for both pollutants, the stepwise and the best random draw models applied to each pollutant have similar $BIC$ scores. Also, the ranked link and LASSO 1SE models are among the best possible scores. In addition, as shown in this chapter, the conclusions point to the clustering method as being that with the best emission estimations for different scales, despite its computational cost.

| | Static/static | | Static/dynamic | |
|---|---|---|---|---|
| VARIABLES → / MODELS ↓ | $CO_2$ | $NO_X$ | $CO_2$ | $NO_X$ |
| **Reference** | 5469 | 3880 | 128111 | 90488 |
| **LASSO 1SE** | 5341 | 3885 | 115976 | 78197 |
| **Random ID 2** | - | 3817 | - | - |
| **Random ID 3** | 5338 | - | - | - |
| **Random ID 7** | - | - | - | 78917 |
| **Random ID 14** | 5363 | - | - | - |
| **Random ID 15** | - | 3742 | - | - |
| **Random ID 16** | - | - | 116580 | - |
| **Random ID 19** | - | - | - | 78499 |
| **Random ID 20** | - | 3766 | 116010 | - |
| **Random ID 26** | - | - | 116779 | - |
| **Random ID 27** | 5401 | - | - | 78241 |
| **4 clusters** | 5382 | 3751 | 117414 | 79275 |
| **6 clusters** | 5343 | 3854 | 117375 | 79048 |
| **Ranked links** | 5345 | **3740** | 115989 | 78118 |
| **Stepwise** | **5328** | 3746 | **115871** | **78034** |

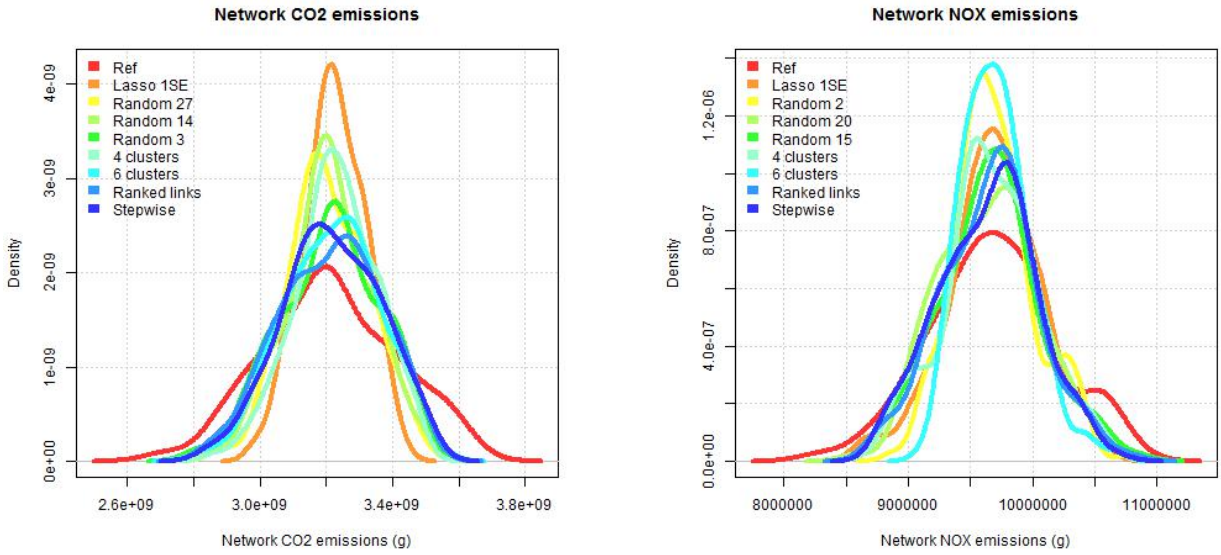
Tab. 4.1 – *The Bayesian Information Criterion for the network emission estimation models.*

### 4.4.1 Static/static dataset

The densities of estimated values of all the methods were compared for the static/static dataset. Figure 4.32 shows the density of the daily estimated $CO_2$ and $NO_x$ emission values.

For the $CO_2$ emissions, the estimated values are around the mean value of the reference. The Random 27 model, Random 14 model and the model built when the network was partitioned into 4 clusters follow the same trend as LASSO but with lower peak density and slightly more dispersed values. The Random 3 (i.e. the best case of the random draws), the "6-cluster" model, the ranked links and the stepwise models have distribution shapes closer to the reference values. The error distributions from the estimated $CO_2$ emission values were compared. Figure 4.33 compares the estimation errors of different methods for $CO_2$ while figure 4.34 focuses on $NO_x$ emissions.

As can be seen in the figure 4.33, all the models studied in this chapter are compared. The sizes of the models derived from each method are shown in parentheses next to the name of the method. The method that selected the fewest links of the network was the stepwise method with only 7 links. When comparing the error distributions, network partitioning obtained the smallest error distributions and the fewest outliers versus the other methods, especially the "6 cluster" model which had 10 random




(a) Density of estimated daily network $CO_2$ emission values.

(b) Density of estimated daily network $NO_x$ emission values.

Fig. 4.32 – *Comparison of the daily emissions values of emissions estimated by the proposed methods.*

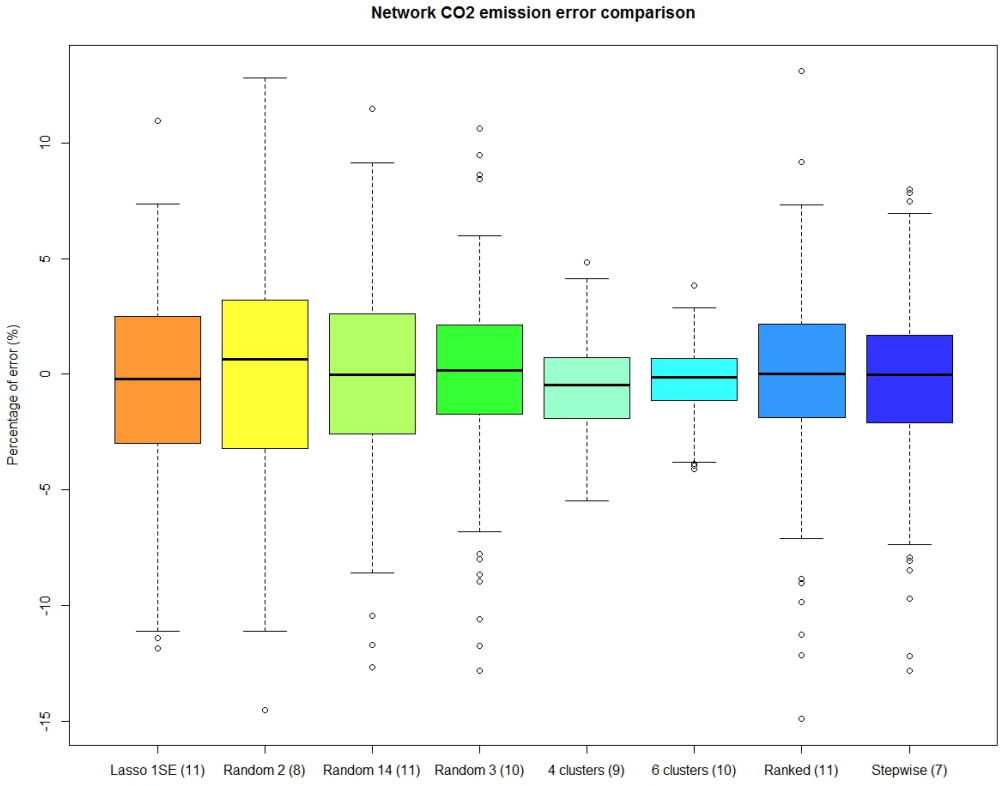



Fig. 4.33 – *Error comparison between the estimated values of all models to estimate the daily $CO_2$ emissions.*

selected links and an error distribution lower than ±4%. Considering that the links in this model were selected randomly, the better goodness of the fit can be explained by the fact that the clusters were defined based on the similarity between the links inside each cluster, which means that any of the links inside it can represent the group to which they belong in the model (i.e. link representativeness).

For the $NO_x$ emission densities, the Random 2 and the "6-clusters" models had similar peak density distributions, while the other models presented similar density distributions. All of them had much larger peak distributions than the reference values. Figure 4.34 compares the error distribution values of all the methods applied up to now.

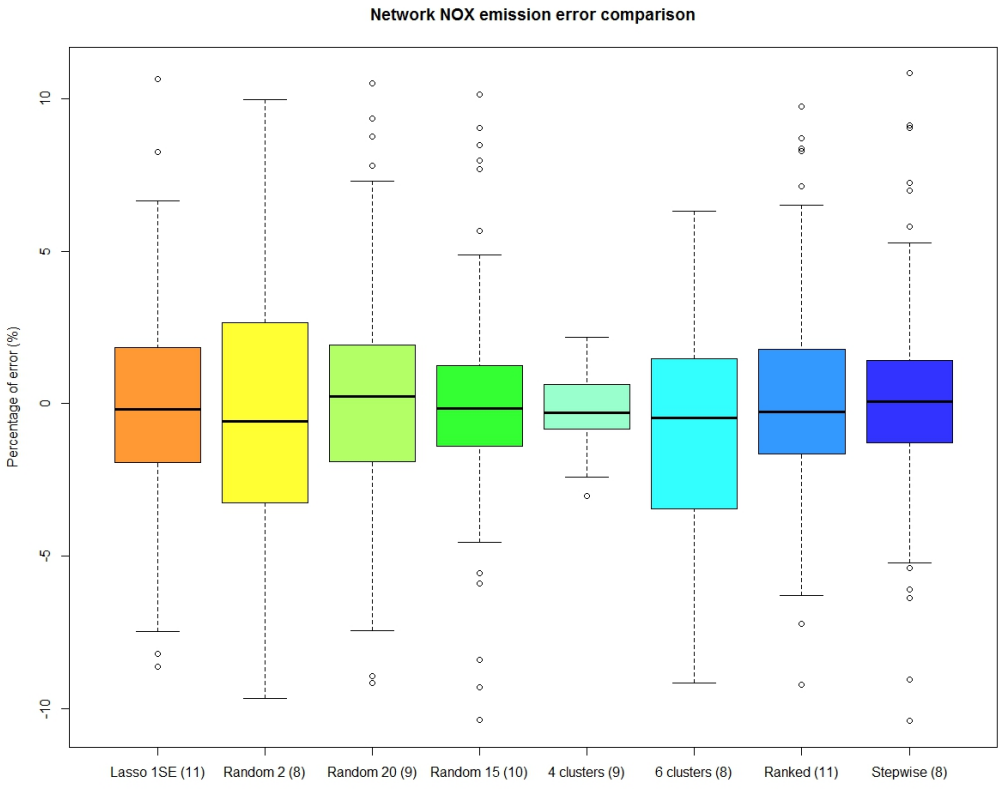


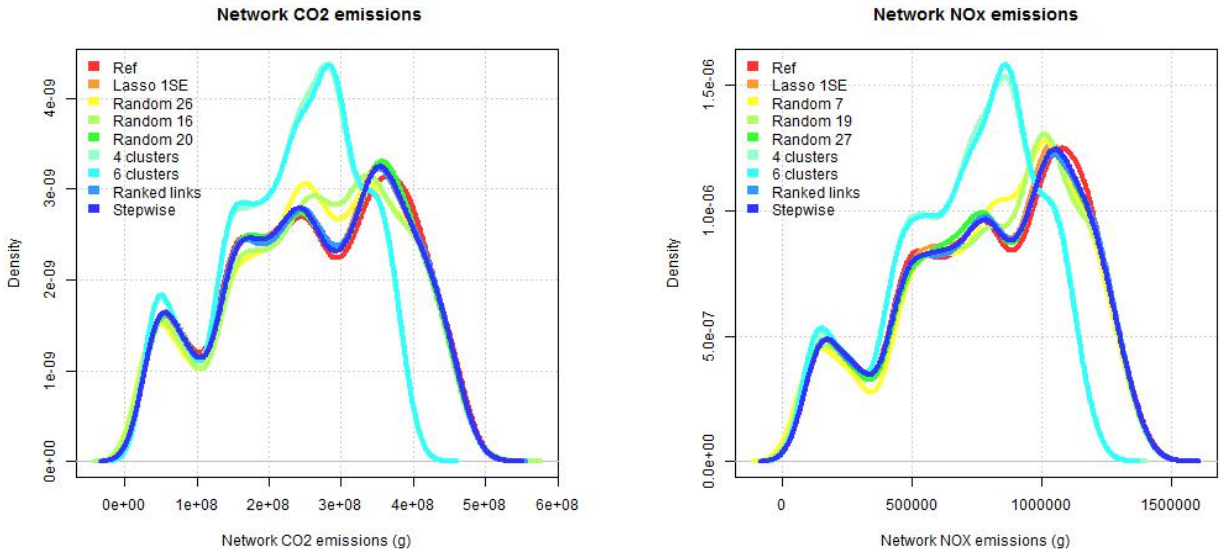

Fig. 4.34 – *Comparison of errors between the estimated values of all the models used to estimate the daily $NO_x$ emissions.*

The Random 2, "6-cluster" and stepwise models were those that selected the fewest links to estimate daily $NO_x$ emissions, with only 8 links. The stepwise model had the least dispersed estimation error, but it had several outliers compared to the other two. Regarding the estimation of $CO_2$, the best model was obtained by network partitioning. The network divided into 4 clusters resulted in a model with 9 links and an estimated error lower than ±3%. As said before, the network partitioned into 6 clusters did not have same level of error because not all of the clusters were represented in the model.

Regarding the daily emissions, all the methods are capable of building models with a reasonable level of error. Network clustering using the "snake" methodology allied with random selection within the clusters gave the best estimation of both pollutant emissions. Regarding $CO_2$ emissions, partitioning the network into 6 clusters estimated the daily $CO_2$ emissions better using only 10 links of the network and a model was obtained that estimated the network emissions with an error lower than ±4%. For the $NO_x$ emissions, 4-cluster partitioning was that which best estimated daily emissions at network level. The estimation error was lower than ±3% with only 9 links in the model

### 4.4.2 Static/dynamic dataset

The densities of the $CO_2$ and $NO_x$ emission values estimated by the models are compared in figure 4.35.




(a) Density of the estimated daily network $CO_2$ emission values.

(b) Density of the estimated daily network $NO_x$ emissions values.

Fig. 4.35 – *Comparison of the density of the daily emission values estimated by the methods proposed.*

Considering the $CO_2$ densities observed in figure 4.35 ($a$), most of the models have a shape similar to the reference values, which was not the case of the models fitted using the clustering method. Likewise with the estimated $NO_x$ values. The density distributions that corresponded most to the reference values were obtained with the LASSO, ranked and stepwise models.

When the error distributions are compared, all the models contained outliers, which could lead to large estimation errors. The figure showing the distribution of outliers is presented in the appendix J.1. Figure 4.36 shows the distribution of errors from the $CO_2$ values estimated by all the models. The same figure containing the outliers is shown in the appendix J.1.

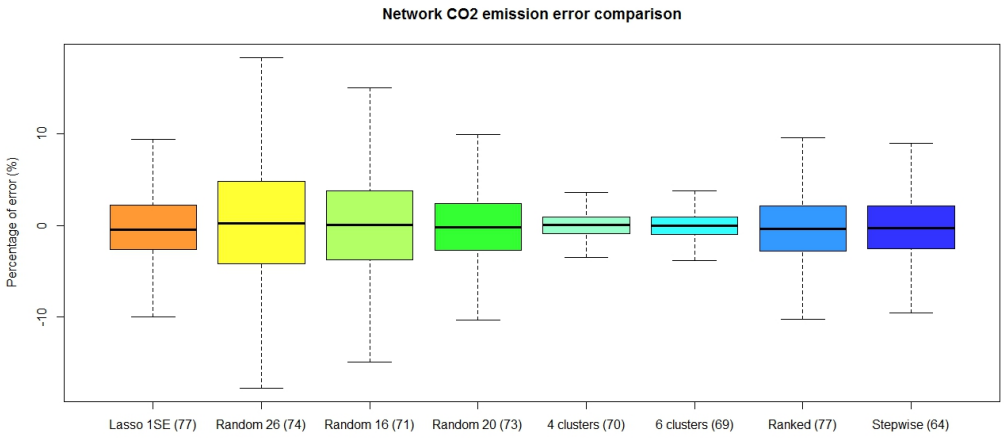



Fig. 4.36 – *Comparison of errors between the estimated values of all the models used to estimate the daily $CO_2$ emissions.*

Considering the outliers of all the models, the network clustering method reduced the number of outliers and vastly reduced their errors. All the other methods had around 6% and 8% of the data as outliers with errors over ±50%. The clustering method reduced the proportion of outliers from 6%

to 2% in the data with ±7.5% fewer errors. Also, the models obtained using the clustering method are those that estimated the emissions with the fewest errors. In both cases, the estimation error was below ±4%, considering that up to 50% of the data had fewer than ±1.0% estimation errors. The difference between them is the number of selected links: 70 links in the network divided into 4 clusters and 69 links for the network divided into 6 clusters. In addition, the model with the network partitioned into 6 clusters had a few outliers compared to the 4-cluster network.

Note that LASSO, the best random draw (Random 20), the ranked links and the stepwise model have the same level of estimation error, with different sets of links and model sizes. With only 64 links the stepwise model had the fewest selected links.

The same analysis was performed for $NO_x$ emissions. Figure 4.37 shows the estimation error on $NO_x$ emissions of all the models studied. The same figure taking into account the outlier distributions is provided in appendix J.1.

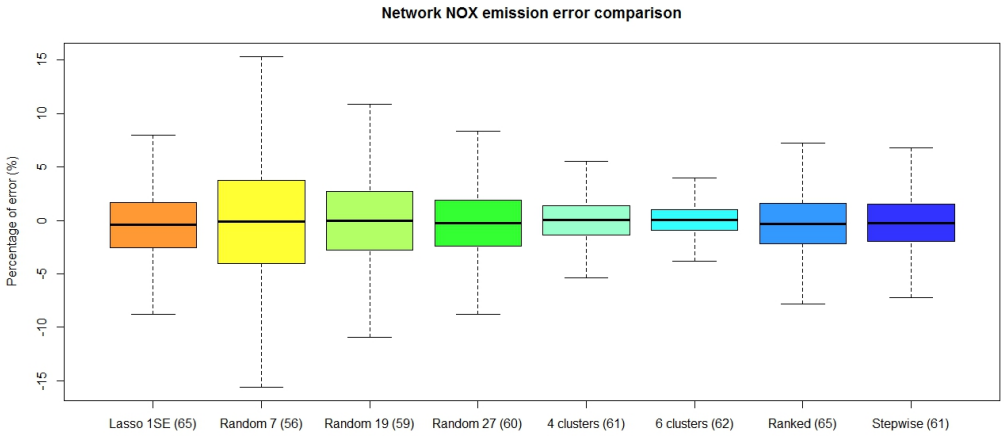



Fig. 4.37 – *Comparison of errors between the values of all the models used to estimate daily $NO_x$ emissions.*

The outliers in $NO_x$ emissions present the same trend as those for the $CO_2$ emissions. In all the models except the clustering models, the outliers represented between 7% and 9% of the data and reached values over -80% and over +40%. The random selection into clusters has only 0.7% of the data as outliers in the "4-cluster" case and 3.7% in the "6 cluster" case. In both cases, the outliers reach errors smaller than ±15%. Also, the error distributions are much smaller than those of the other methods, mainly in the network divided into 6 clusters with 62 links in the model, for which the percentage error is lower ±4%. Regarding $CO_2$ emissions, the LASSO, the Random 27, the ranked and stepwise models obtained similar error distributions with different model sizes, of which the smallest was the Random 27 with 60 links inside.

## 4.5 Case of application

The City of Paris adopted a standard license type Open Database for various aspects of development and control in Paris. The major public roads of Paris are equipped with vehicle metering stations (i.e. loops) for traffic regulation purposes and to provide information for users. We retrieved the geographical location of the sensors from the database available at (Mairie de Paris, 2015). Using their geographical location, it was possible to identify the link sections of the road system built in

the traffic simulator. The idea is to use the simulated data from these identified links with loops to estimate emissions at network level. The methodology is the same as the others; a model is built using the linear regression of the data obtained from the links with loops, to estimate the total emissions of the entire network. The structure of the data is the same as that used in the previous section: the static/static dataset considers the daily values of each link and the static/dynamic dataset the values of 15 minute time periods.

According to the database available in (Mairie de Paris, 2015) and our road system integrated in the traffic simulator, the $6^{\text{th}}$ district of Paris has 45 links equipped with loops placed on the major corridors. These loops correspond to 20% of the network links. This percentage is higher than the selection rate of pollutants studied in the static/static dataset and lower when considering the static/dynamic dataset. Taking this into account, the aim is to see whether it is possible with these 45 links to reduce the associated network emission estimation errors in the static/static dataset, and also reduce the number of links in the static/dynamic models. Figure 4.38 shows the links with sensors in the $6^{\text{th}}$ district of Paris.

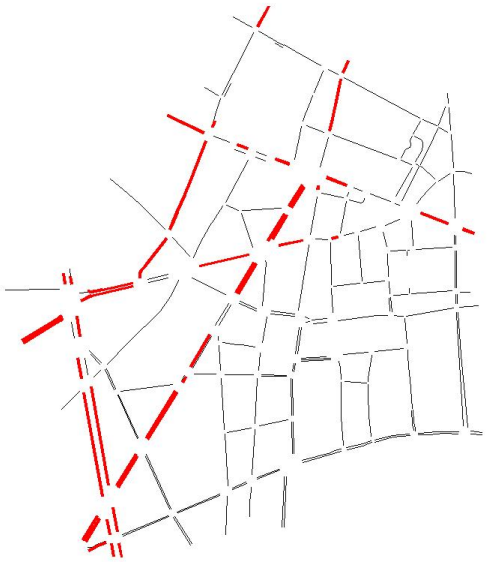
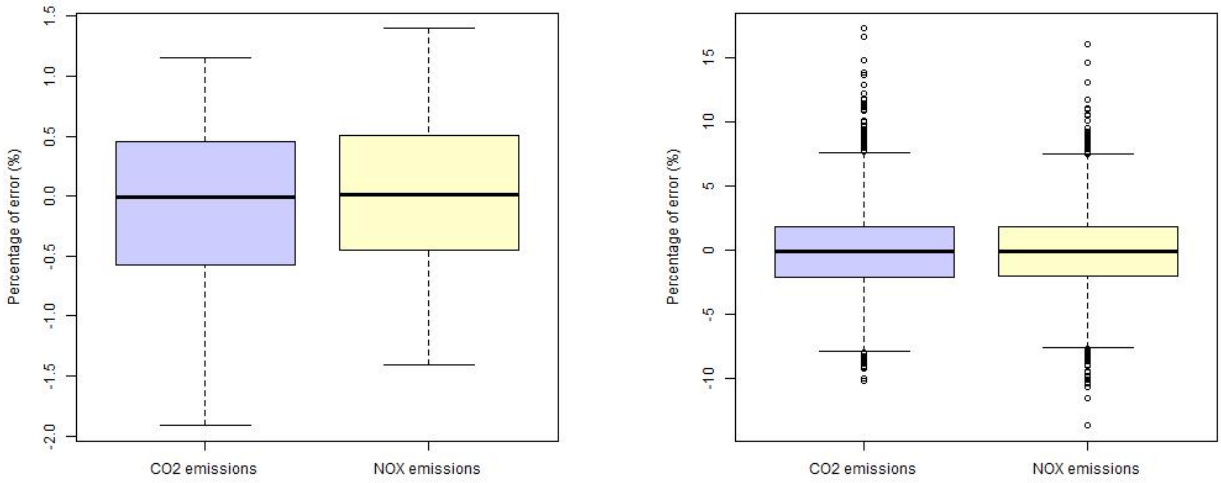


Fig. 4.38 – *Links equipped with traffic sensors in the $6^{th}$ district of Paris.*

These links correspond to about 44% of the total emissions according to the simulated data. A linear regression is applied to the data from these links considering the $CO_2$ and $NO_x$ pollutants. The associated pollutant estimation errors in both datasets are shown in figure 4.39.

Using the links with a traffic sensor can significantly reduce the associated estimation errors of the static/static dataset. Furthermore, both pollutants do not present outliers, and all the observations can be considered by the model with a confidence interval of 5%. The $CO_2$ emission estimations presented errors between -1.9% and 1.2%. Considering this distribution, 50% of the estimated values have errors between -0.6% and 0.5% with a median value of -0.01%. The estimated $NO_x$ emission values present slightly more errors when compared to the $CO_2$ values, but they are still smaller compared to all the values estimated previously by the models studied in this section and by LASSO. The estimated $NO_x$ values have errors between $\pm$1.4%. Considering all the error values, 50% of them are between -0.5% and 0.5%.




(a) Error distributions in the static/static datasets.

(b) Error distributions in the static/dynamic datasets.

Fig. 4.39 – *Error distribution of the network emission estimation.*

Using the daily values of these 45 links equipped with traffic sensors, the total emissions of the entire network can be estimated from these links considering a percentage error lower than 2%. The models are 4 times bigger than those of LASSO and the stepwise selection, but with a considerably smaller error. Figure 4.39($b$) shows the associated errors of the emission estimations in this dataset.

The static/dynamic datasets represent the 15 minute traffic data, and also more varied network data. This type of data needed at least 25% of the links to estimate the network emissions with a reasonable error distribution of around ±10%. The number of equipped links in the $6^{\text{th}}$ district of Paris is smaller than the number of links selected by the LASSO and stepwise methods. Generally, with fewer links compared to the previous sampling methods, the errors are less dispersed and the number of outliers is smaller. Considering both pollutant emissions, the errors are dispersed between -7.9% and 7.6%, and 50% of these errors are between -2.0% and 1.9%. The outliers represent 2.8% in $CO_2$ and 2.4% of the data in $NO_x$. This is the smallest rate compared to the other models already studied.

As concluded before, in this case study, no single set of links can be used to estimate the total emissions of the network. Besides, many links of the network are interchangeable. Comparing all the models analyzed until now and taking into account their error distribution, the links currently equipping the Paris network have the smallest error distribution considering 15 minutes as a time period for traffic data and for daily values. All of them are placed on a major corridor on which traffic demand is intense.

## 4.6 Conclusion of the chapter

In this chapter other sampling methods were proposed for comparison with the optimal model from LASSO for each pollutant and dataset. The aim was to propose the simplest sampling method to compare the emission estimation results at the network level. The methods were: (i) random selection over the network and (ii) link emissions ranking from the most pollutant to the least pollutant (both were assumed to select the same number of links as LASSO did in chapter 3 for each pollutant and

dataset), (iii) network partitioning using the "snake" method, and (iv) stepwise selection.

Regarding computation time, the results were obtained very quickly for (i) and (ii). The stepwise method had shortest computation time, about 10 seconds versus the LASSO algorithm that required around 30 seconds for each pollutant and dataset. Network partitioning took around 60 seconds to divide the network into clusters. Although it was the methodology that provided the best results in terms of estimation, this computation time could be an obstacle if it is applied to bigger networks.

The estimations of all the methods presented were the same as or better than those obtained with LASSO. All the models had their own set of selected links. Most of models were smaller than LASSO and had the same level of estimation error. Of all the models, network partitioning gave the models that best fitted the network emissions to both pollutants and datasets based on their estimation errors. The models were also evaluated using the $BIC$ scores for both pollutants and datasets. The stepwise, LASSO 1SE and ranked links models obtained the lowest scores. Regarding the static/static dataset, the best random draw for each pollutant also obtained one of the lowest scores.

For policy makers, network clustering is an alternative for improving the prediction and estimation of network values using only a few links representing all the cluster of the network. Also, the simplest methods such as ranking the most pollutant links on the network can give network emission estimations for any spatial-temporal range using the same quantity of link information and obtain an estimation with the same level of error as LASSO.

Many studies are required to confirm the results of this chapter but this work shows that simple solutions can be applied to estimate network emissions with few data.

# 5

# Effect of traffic data aggregation on emission estimations

The application of information technology to traffic management has made it possible to collect huge volumes of data on urban traffic. The states, characteristics and attributed values of populations, economic activities and transportation facilities change over time, giving a dynamic dimension to these factors (Ichikawa et al., 2002). In addition, the inherent dynamism of mobility directly influences the spatial and temporal variability of urban travel. Having access to continuous knowledge of urban dynamics plays a fundamental role in the efficiency of strategies and the management of urban traffic, particularly regarding the reduction of traffic pollution (Stathopoulos and Karlaftis, 2001). As emissions are a complex function of many variables, impacts and solutions are commonly evaluated using multidisciplinary combinations of transport and emissions at different scales, ranging from local management at road level to entire urban transport networks.

Data processing methods have helped to facilitate understanding and the development of many applications for urban traffic operations and planning. However, the conversion of large volumes of data into useful information for public managers requires pre-processing, modeling and prior post-treatment, corresponding to a series of analyses to obtain a clearer and more legible view of the results. This complexity requires the systematization of data collection and processing, but in the case where information is either inadequate or insufficient, it is necessary to use models to simplify and adjust methods to the availability of data. To simplify the process and also shorten processing time, planners tend to gather spatial and temporal traffic data to obtain a global view of the situation. Nonetheless, this global view cannot represent what happens locally. This procedure was observed in particular by (Barth et al., 2000) when estimating $CO_2$. Spatial and temporal data aggregation can lead to gaps in knowledge depending on the level of collection. Considering the traffic variables commonly used to estimate emissions, they cannot be exactly the same as a function of spatial and temporal aggregation if they do not share the characteristic of linear behavior. This presupposes that the emission estimations derived from these traffic variables are biased depending on the scale. This is related to the function result when averaging non linear process, *i.e.* $f(\bar{v_i}) \neq \overline{f(v_i)}$ Furthermore, when focusing only on the emission functions of COPERT IV (Gkatzoflias et al., 2012) and (Ntziachristos and Samaras, 2014), it can be seen that their speed-factor curves have a parabolic shape, indicating that they are non-linear.

The aim of this section is examine the influence of spatial and temporal aggregation on the traffic variables used to estimate pollutant emissions of COPERT functions. Simulated data of the $6^{\text{th}}$ district of Paris were used to explore these relations. Initially, two scales of spatial and temporal aggregation were proposed. Regarding spatial aggregation, the local scale is defined by traffic variables at link level

while the global scale takes into account the entire district. The objective of both scales is to estimate emissions at network level. Considering temporal aggregation, traffic data can be recovered directly by sensors every 15 minutes and also be grouped for daily traffic. These two scales applied to the case of Paris allow carrying out a sensitivity analysis to determine to what extent the possible range of these input variables influences model outcomes (i.e. $CO_2$ and $NO_x$ emissions for road links) and hence accuracy. The accuracy of emission estimations can be affected by the errors associated with the emission model itself (i.e. emission factors), and the errors associated with the input variables in the emission model. Road level studies commonly use measured input data for key variables such as vehicle kilometers traveled (i.e. traffic volume multiplied with road length), mean speeds and fleet mix composition to validate the method as presented in (Pierson et al., 1996) and (Mukherjee and Viswanathan, 2001). Consequently, these studies tend to quantify the errors associated only with the emission model, so they validate the methodology applied but do not directly assess the accuracy of coupled traffic-emission estimations and in particular the scale at which the coupling is performed. Similarly, extending the area of validation to an area larger than a road makes it possible to study and quantify associated errors, since the estimations are based on the combination of traffic and emission models. The purpose is to quantify the errors and highlight that a bias can be identified. This shows that such calculations lack consistency, in particular when traffic dynamics and congestions are taken into account correctly. Afterwards, the COPERT functions are analyzed to understand how they were constructed and how they can be used correctly to estimate emissions and why the "scaling" bias appears.

## 5.1 Local versus global emission calculations

### 5.1.1 Data input

The simulated data described in chapter 2 were used as the basis of this study. As in the previous chapters, the coupling emission functions from COPERT and traffic data for estimating emissions at network level were used. To this end, two main coupling scales are defined: (i) emissions are calculated for all the links based on travel distances and the mean speeds for each period of time and then summed to determine the daily network emissions; and (ii) traffic variables are first aggregated at network level and then the emissions are derived based on the aggregated values. In the first method, called local scale, the traffic variables are recovered every 15 minutes at each link and then the emissions are calculated directly with these values. Afterwards, all the emission values at link level are collected to estimate the total values of daily emissions at network level. The second calculation method, called global scale, starts by grouping all the traffic variable values from the link level to the network level after which the emissions are calculated based on these values. These two spatial-temporal scales are identified to compare the emission calculation results derived from the same traffic data but aggregated differently. For each scale, the impact on the model output is then evaluated by testing the influence of the aggregation on multiple simulations with different settings.

The emissions using COPERT methodology were determined by three main variables, namely traveled distance (i.e. traffic volume multiplied with road length), mean speed and traffic composition. The amount of travel had a considerable effect on emission estimations, particularly because any error in the total distances traveled by cars is propagated proportionally in the emission estimations. Compared to the other variables such as mean speeds and fleet composition, accurate traveled distances on streets are easy to obtain as traffic count data measured at various points (e.g. sensor detection,

cameras, manual counting survey) in road networks. For the dynamic microscopic model, all the exact values of the distances traveled by the cars were recovered from each link and considered as linear variables. There were no differences between either of the scales proposed.

Inaccurate speed inputs can have a big effect on estimated emissions (Smit et al., 2008). The same author showed that a change in the joint distribution of mean speed can produce a difference of up to 9% in $CO_2$ and $NO_x$ estimations using the factor curves of COPERT IV. (Chatterjee et al., 1997) conducted a sensitivity analysis on the average speed emission model MOBILE5 and showed that an error of 5 km/h in the value of mean speed used as an input of an emission model for a freeway caused a 42% difference in the estimation of $CO$ emissions due to the strongly non-linear relationship between the emission factor curves and mean speeds. Here, two scales were defined and applied, namely local and global. The differences in mean speed values on each scale and how they could affect emission estimations were explored.

### 5.1.2 Emission quantification

To refresh the reader's memory, the equation employed to assess pollutant emissions using emission factor curves and traffic data is shown in 5.1.

$$e_p = d_i \times f(\bar{v_i}) \tag{5.1}$$

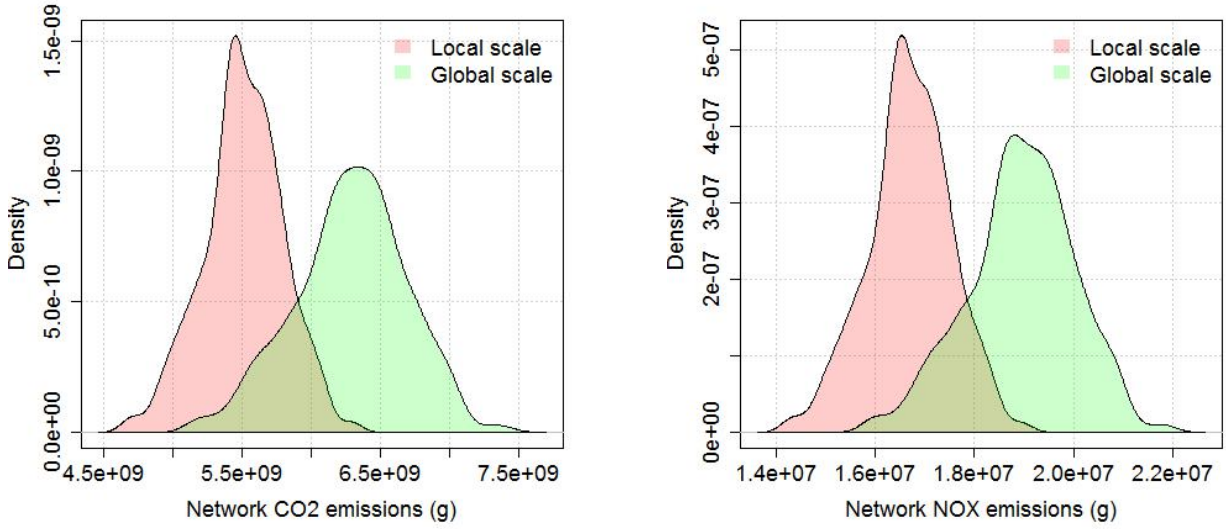
where:

- $e_p$ is the pollutant emission in g/km;
- $d_i$ is the traveled distance in the spatial element $i$ (link);
- $f(\bar{v_i})$ is the pollutant emission factor associated with a passenger car fleet determined by average speed in $i$;
- $\bar{v_i}$ is the average speed in $i$.

Considering the microscopic traffic data, the emission factors for hot exhaust emissions and the French fleet composition, the $CO_2$ and $NO_x$ emissions were calculated for both the scales proposed, local and global, and they are shown in figure 5.1.

Figure 5.1 shows the daily pollutant emissions of the network (i.e. 400 measurements under different traffic conditions). As can be seen, the emission quantification results on both scales are completely different. For both pollutants, the emissions calculated with the global scale data have higher mean values, and the data are more dispersed than those of the local scales. Analysis of the gap in percentage between both scales, as shown in figure 5.2, the difference between daily emissions is on average around 14%.

Regarding the pollutant emissions, in this case carbon dioxide, the disparity can be between 10% and 18%, and up to 50% of these emission values are between 13% and 16% of the disparity. The disparity of the nitrogen oxides is less scatted compared to CO2. It is between 11% and 16% considering that 50% of the emission values are between 13% and 14.5% of the disparity. Comparing both pollutants, $CO_2$ has larger and more dispersed disparities than $NO_x$, due to the shape of the emission functions. For the same mean speeds used for both scales, $CO_2$ appears to have larger gaps than $NO_x$.




(a) Density distribution of $CO_2$ emissions.

(b) Density distribution of $NO_x$ emissions.

Fig. 5.1 – *Density distributions of network emissions using the local and global approaches.*

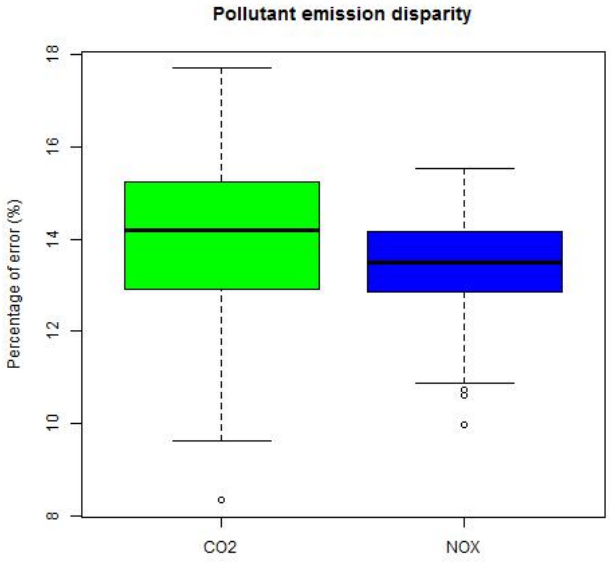
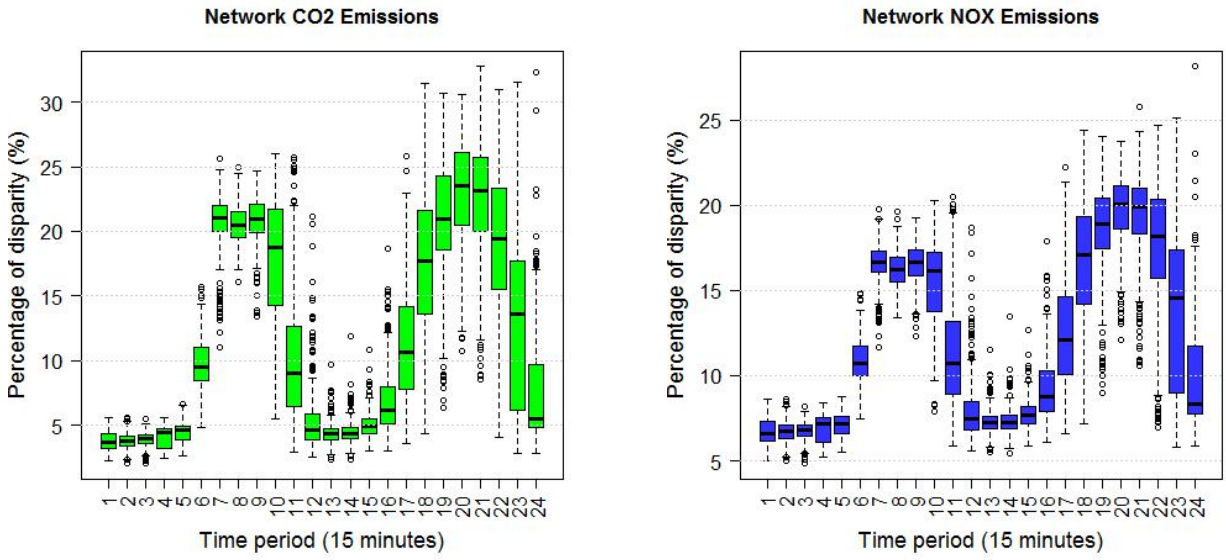



Fig. 5.2 – *Percentage of disparity between the daily emission estimations for each pollutant.*

The densities presented show the results of daily emission estimations considering spatial and temporal aggregation at the same time. To assess the influence of the traffic states on this bias between both scales, the disparities were analyzed by time period. Figure 5.3 shows the disparities between the amount of emissions estimated by the global and local approaches by time period.

Two traffic states stood out in all the simulations: free flow and congestion states. The time periods defined as congested states are between periods 7 and 13, and 20 and 21. In these periods no trends are apparent and the disparities are varied. For the first periods of congestion, most daily emissions have a gap between 15% and 20%. When the level of congestion starts to decrease, the disparities also decrease on average but are more spread out compared to the others, as observed for time period 11. The transition between congested and free flow states is highlighted for periods 13 and




(a) The $CO_2$ emission disparities by time period.

(b) $NO_x$ emission disparities by time period.

Fig. 5.3 – *Disparities on emission estimations between scales by time period.*

14. The disparities are almost constant for all the 400 simulation days, and are around 10% for $CO_2$ and about 8.5% for $NO_x$. For periods 20 and 21, they represent the largest gaps of all the periods, for $CO_2$ emissions period 20 is larger than 21 and represents a difference between 16% and more than 30%, and for $NO_x$ both periods are similar and present a gap between the scales of slightly less than 15% and 25%. To conclude, gaps can be found in all the traffic states and can vary between 5% and 25% on average. The congested states have larger disparities and these deviations can reach about 23%; these disparities are also more dispersed in the congested states. For the free-flow conditions, the differences between both scales are smaller compared to the congestion periods but not neglected. They vary around 4% for the CO2 and 7% for the $NO_x$ emissions. Considering the pollutants, $CO_2$ has larger gaps in general compared to $NO_x$. This is normal because when the traffic conditions are homogeneous (free-flow), then the difference between both methods of emission quantification are low by definition.

These results show that using the COPERT emission functions at various spatial-temporal scales induces a bias depending on the aggregation of traffic data when considering traffic dynamics, *i.e.* the time variation of the speed. For the same data, which represent more than a year of daily traffic and various demand levels, aggregation at the local scale (i.e. link) and the global scale (i.e. network) provides different quantifications of pollutant emissions. The next section describes the analysis of the biases is investigated at another scale. We are going to aggregated driving cycles with different time periods to see the difference in emission estimation. This will help to better the bias that cames from ythe averaging of speed values that is not scalable.

## 5.2 The analysis of the biases

### 5.2.1 Driving cycles

Driving cycles are essential to quantify the impacts and of road transport and for certification, because they are the main link between kinematics and driving conditions. Vehicle driving cycles are a series

of data points representing the speed of a vehicle versus time. The cycle reflects the real working conditions of a vehicle or engine under specific traffic conditions, thus it provides a reasonable evaluation of the emission function of the vehicle. (Andre et al., 1995) summarized driving cycle method by the following steps: (i) data collection; (ii) data segmentation; (iii) cycle construction, and (iv) the evaluation and selection of the final cycle. These steps were used to assimilate driving conditions on a laboratory chassis dynamometer to evaluate fuel consumption, exhaust emissions and emission coefficients (Simanaitis, 1977).

There are many categories of driving cycles and they are classified into two types: modal and transient. The main difference is that modal cycles are a compilation of straight acceleration and constant speed periods and are not representative of real driver behavior, whereas transient cycles involve many speed variations typical of on-road driving conditions. Examples of modal driving cycles are the NEDC (New European driving cycle) and the ECE (also known as the MVEG-A cycle) that are composed of various driving modes of constant acceleration, deceleration and speed. The other type is derived from real driving data and is referred to as the "real world" cycle, such as FTP-75 (DieselNet, 2017) and Artemis, which means they are more dynamic and reflect the more rapid acceleration and deceleration patterns experienced under real road conditions (Nicolas, 2013). All these characteristics of both types of driving cycles can be observed in figures 5.4 and 5.5.

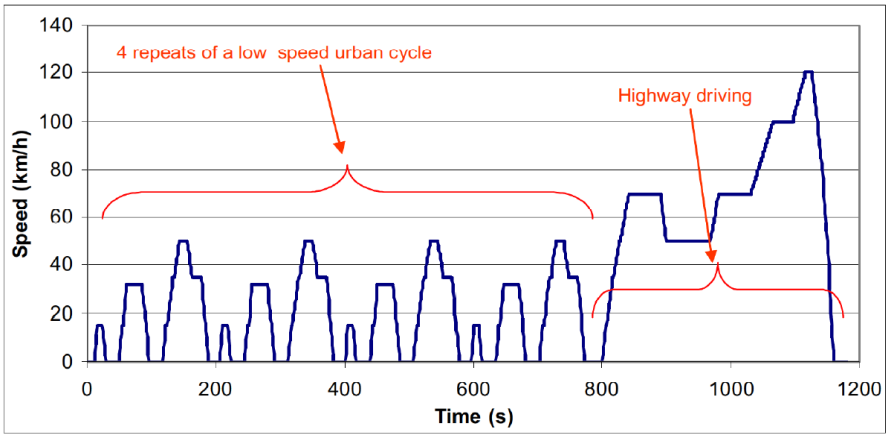



Fig. 5.4 – *Example of the modal driving cycle (NEDC) (Barlow et al., 2009).*

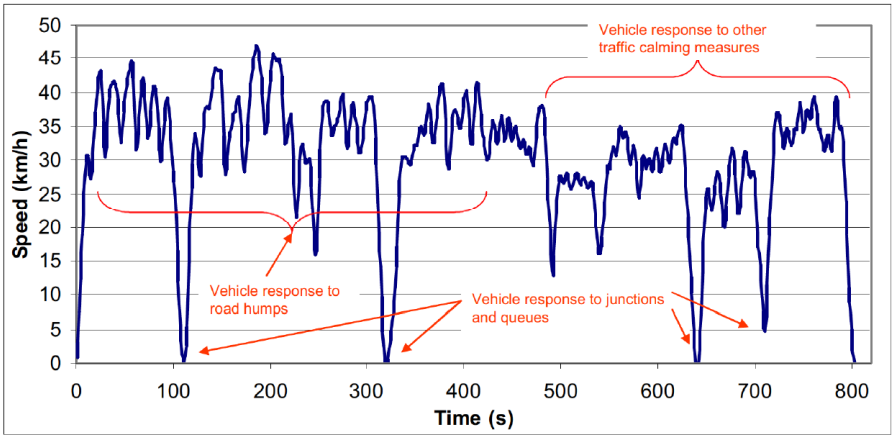



Fig. 5.5 – *Example of a transient driving cycle (Barlow et al., 2009).*

The COPERT emission functions are based on an NEDC driving cycle and concerns have been expressed that this driving cycle is not representative of real-world driving conditions. More studies

have been conducted to develop the functions inside COPERT so that they are closer to real world conditions (Katsis et al., 2012).

The COPERT functions are based on experimental data from European driving cycles, with a series of data points representing vehicle speed versus time that defines the emission factor for each average speed. Each point represents the driving cycle profile of a vehicle belonging to a vehicle category. For each vehicle class, the emission factors are estimated for each pollutant emission according to the weighting of estimated emission factors that belong to an urban fleet mix. The driving cycle procedure requires very extensive databases including different types of car, driver, driving situations, etc. to ensure the representativeness of the population. The idea underlying the driving cycles is to produce a scenario consisting of accelerations, decelerations and frequent stops at constant speed over a period of 20 minutes. The speed at any time during the test must be maintained within a certain tolerance range around a predefined set point.

The cost and the complexity of these experiments mean that, in general, limited samples of vehicles are tested on regulatory cycles which have the advantage of relying on existing data and constitute a reference and basis for comparison. Considering these simplifications of emission functions built on driving cycles, and the fact that they are coupled with traffic models for extrapolation to larger scales compared with the original scale (i.e. driving cycles), the difference in scale can induce a bias in the estimation.

### 5.2.2 Emission gap as a function of data aggregation

Emissions factors are commonly associated with average speed, and researches use the average speed as traffic performance measurement. For each vehicle class, the emission factors are estimated for each pollutant emission according to the weighting of estimated emission factors that belong to an urban fleet mix. The driving cycle procedure requires very extensive databases including different types of car, driver, driving situations, etc. to be representative of the population.

Taking all these considerations into account, the same local and global scale approaches were used to analyze a hypothetical driving cycle. The aim is to observe the bias that can occur when the emissions are estimated over the complete driving cycle (i.e. global scale) or when emissions are first estimated on sub-driving cycle (i.e. local scale) before being aggregated. Figure shows a hypothetical driving cycle.

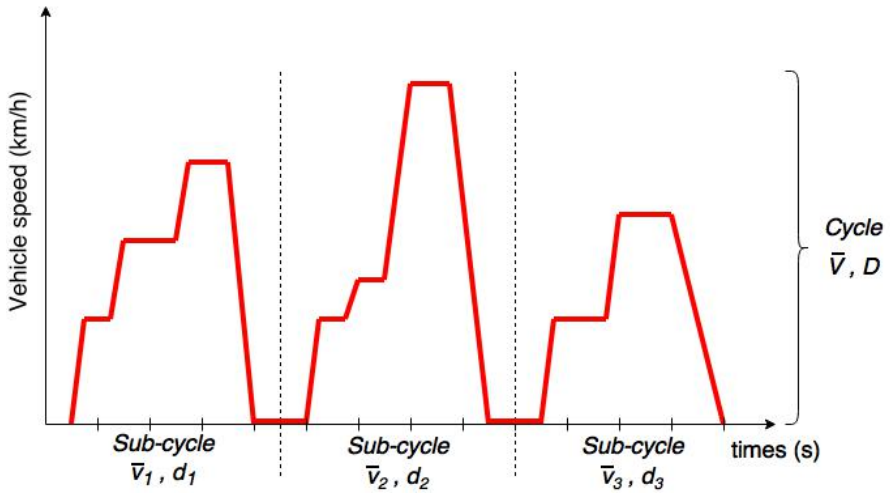



Fig. 5.6 – *Hypothetical driving cycle.*

Figure 5.6 shows an instantaneous speed of a vehicle in a given period of time. In figure 5.6, it is possible to calculate the average speed and the traveled distance of the vehicle by considering the whole cycle (i.e. global approach). Then, the emission calculation is performed using the COPERT functions, as shown below:

$$e_{cycle} = D \times f(\bar{V}) \tag{5.2}$$

where:

- $e_{cycle}$ is the total emissions of the cycle;
- $D$ is the total traveled distance by a vehicle in the cycle;
- $f(\bar{V})$ is the emissions factor function determined by the average speed of the cycle;

The same driving cycle can be divided into 3 sub-cycles with the same length of time, and in each cycle the mean speed ($\bar{V}$) and the traveled distance $D$ are calculated. To quantify the emissions for the entire cycle considering the sub-cycles, the emission calculation is performed as follows:

$$e_{sub-cycle} = d_i \times f(\bar{v_i}) \tag{5.3}$$

Consequently the emissions for the entire cycle are calculated as follows,

$$e_{cycle} = \sum e_{sub-cycle} = \sum_{i=1}^{n} d_i \times f(\bar{v_i}) \tag{5.4}$$

Equation 5.3 shows the emission calculation for each sub-cycle defined using the traffic data which refer to each sub-cycle. Equation 5.4 quantifies the emissions of the sub-cycles to determine the emission for the entire cycle (i.e. local scale approach). Considering that COPERT emission factors can be interpolated by a cubic equation, the emission equations 5.3 and 5.4 can be described as follows:

$$e_{cycle} = D \times (a_0 + a_1\bar{V} + a_2\bar{V^2} + a_3\bar{V^3}) \tag{5.5}$$

and for the local scale,

$$e_{sub-cycle} = a_0 \sum_{i=1} d_i + a_1 \sum_{i=1} d_i\bar{v_i} + a_2 \sum_{i=1} d_i\bar{v_i}^2 + a_3 \sum_{i=1} d_i\bar{v_i}^3 \tag{5.6}$$

The bias function between the two methods can be calculated by the difference between the emission at cycle scale as in 5.5 and the total emissions estimated in n sub-cycles as in equations 5.6, as shown in 5.7 and 5.8.

$$\Delta = e_{cycle} - e_{sub-cycle} \tag{5.7}$$

Consequently,

$$\Delta = a_1 D \times \left(\bar{V} - \sum \frac{d_i}{D}\bar{v_i}\right) + a_2 D \times \left(\bar{V}^2 - \sum \frac{d_i}{D}\bar{v_i}^2\right) + a_3 D \times \left(\bar{V}^3 - \sum \frac{d_i}{D}\bar{v_i}^3\right) \tag{5.8}$$

The emission function is not a scalable function, which means the difference between both scales from the same cycle is considered, the $\Delta$ value is different from zero due to the non linear function.

The more convex the emission functions is, the higher the bias. Note that emissions curves are shown in Chapter 2 figure 2.10.

By taking these considerations into account and extrapolating them to a traffic network, certain probability properties can be applied to define the bias between both scales. The central limit theorem is a statistical theory that gives a sufficiently large sample size for a population with a finite level of variance. The mean of all the samples from the same population is approximately equal to the mean of the population. Furthermore, all the samples follow an approximate normal distribution pattern, regardless of the underlying distribution, with all the variance being approximately equal to the variance of the population divided by the size of each sample.

To better illustrate the definition of the theorem, the sample must contain a large number of observations and each one is generated randomly in a way that does not depend on the values of the other observations. Then, the arithmetic averages of the observed values are computed for each set of observations and this procedure is repeated many times. The theorem distributes the computed values of the average according to the normal distribution.

Keeping in view all these considerations, the emission quantification based on the local level (i.e. links), as defined in 5.4, can be estimated considering that the random variables which describe the links of the network are:

$$X = d_1, d_2, d_3, \ldots, d_n \tag{5.9}$$

$$Y = f(\bar{v_1}), f(\bar{v_2}), f(\bar{v_3}), \ldots, f(\bar{v_n}) \tag{5.10}$$

Where $X$ describes the total traveled distances on each link and $Y$ describes the emission factor that corresponds to the average speed of each link. For a sufficiently large n and considering that the random variables are independent and identically distributed, using the central limit theorem the local emissions can be described as 5.11.

$$e_{local} = n \times E(X, Y) \tag{5.11}$$

Developing the equation,

$$e_{local} = n \times E(X)E(Y) + n \times \mathrm{Cov}(X, Y) \tag{5.12}$$

So,

$$e_{local} = \sum d_i \times \frac{\sum f(\bar{v_i})}{n} + n \times \mathrm{Cov}(d_i, f(\bar{v_i})) \tag{5.13}$$

Equation 5.13 describes how the network emission can be calculated using the local data at each link to estimate the daily pollutant emission values. The bias of the network emissions calculated from the global scale and the local scale, as defined in 5.7, can be estimated as follows:

$$\Delta = D\left(f(\bar{V}) - \frac{\sum_i f(\bar{v_i})}{n}\right) - n \times \mathrm{Cov}\left(d_i, f(\bar{v_i})\right) \tag{5.14}$$

Considering that $\bar{V}$ is the weighted harmonic mean of $d_i$ as shown in 5.15, which represents the average speed at the global scale and in 5.16 the local mean speed $\tilde{V}$ is represented by the weighted

arithmetic mean of $d_i$ considering the properties of the theorem, then

$$\bar{V} = \frac{D}{\sum \frac{d_i}{v_i}} \tag{5.15}$$

and,

$$\tilde{V} = \frac{\sum \bar{v}_i}{n} \tag{5.16}$$

To preserve convexity,

$$\Delta = D\left(f(\bar{V}) - f(\tilde{V})\right) + D\left(f(\tilde{V}) - \frac{\sum_i f(\bar{v}_i)}{n}\right) - n \times \mathrm{Cov}(d_i, f(\bar{v}_i)) \tag{5.17}$$

As can be seen in 5.17, the bias between the global and local scales can be estimated by three terms: the first defines the gap between the emission functions based on the local and the global scales. The second term corresponds to the convexity of the function, and the last term that corresponds to the correlations between the variables, i.e. $d_i$ and $f(\bar{v}_i)$. This highlights the fact that the bias between the scales strongly depends on the mean speed. Because of its non-linear nature, the mean speed will not be the same in any spatial and temporal aggregation of the same population.

The first term refers to the difference between the mean speeds defined in 5.15 and 5.16. To illustrate this, both mean speeds are calculated by time period using the same simulated traffic data. The comparison by time period is shown in figure 5.7. Note that this bias can be calculated if both mean speeds values are known at large scale.

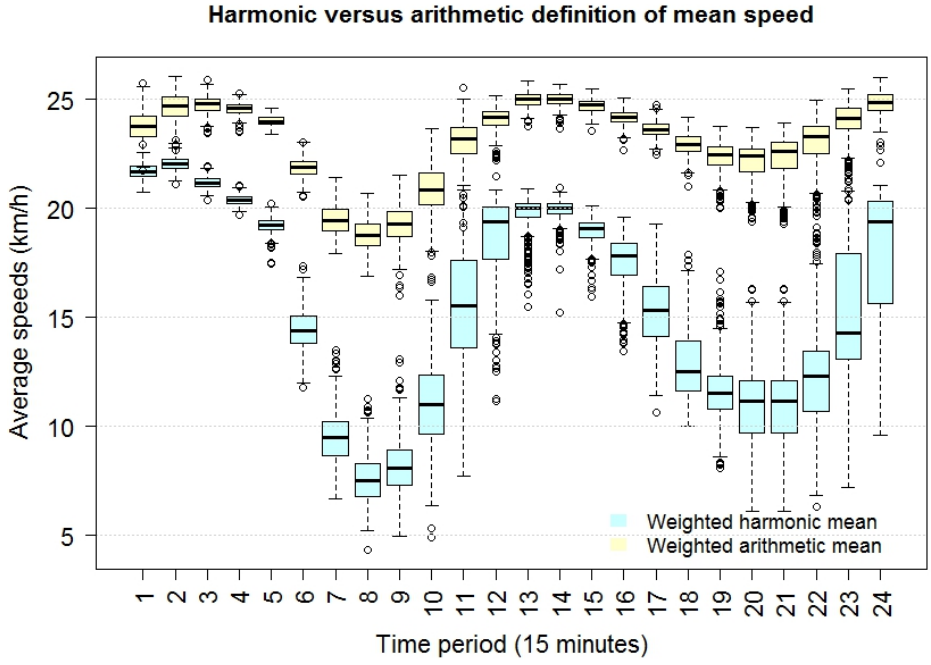
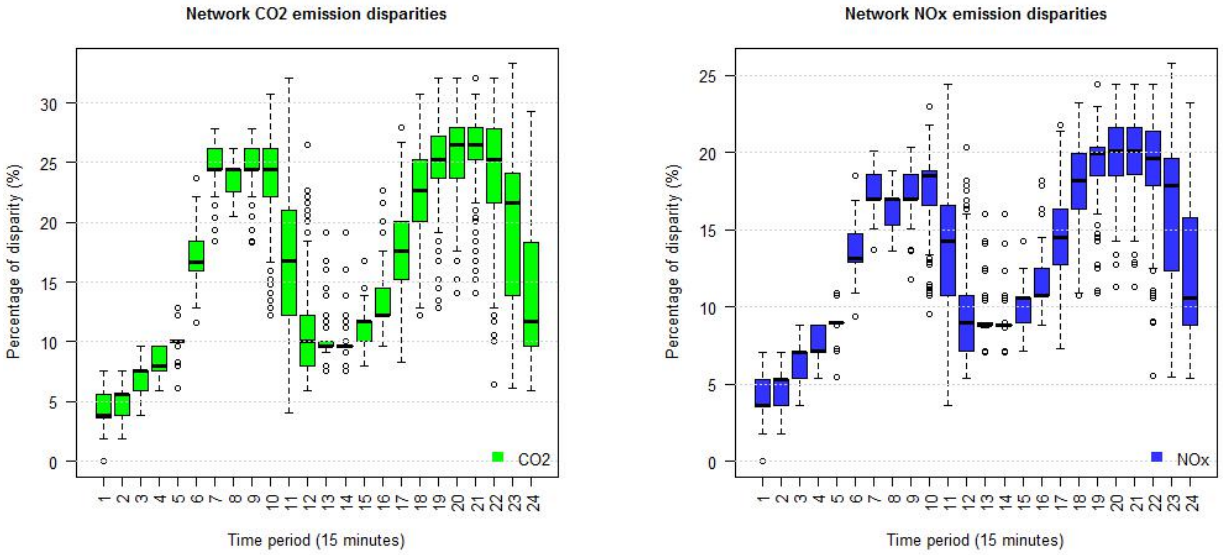



Fig. 5.7 – *Comparisons between the weighted and arithmetic mean speeds by time period.*

As can be observed, when considering all the simulations with different traffic dynamics in the network, the arithmetic definition of the mean speed is higher than the harmonic mean speed. Also, the distribution values are less dispersed, especially in the periods when the network is considered in free flow state. The gap between both mean speeds increases when the network is congested, as shown

in the periods between 7 and 13, 21 and 22 where the values are more dispersed. Another point is that the arithmetical mean speed tends to reduce the variability of the data when compared with the harmonic definition.

Using the factor-speed curve of each pollutant, the emission factors are calculated for each mean speed definition. The gaps between them were calculated and are shown for each pollutant in figure 5.8.




(a) $CO_2$ emission factor disparities by time period.

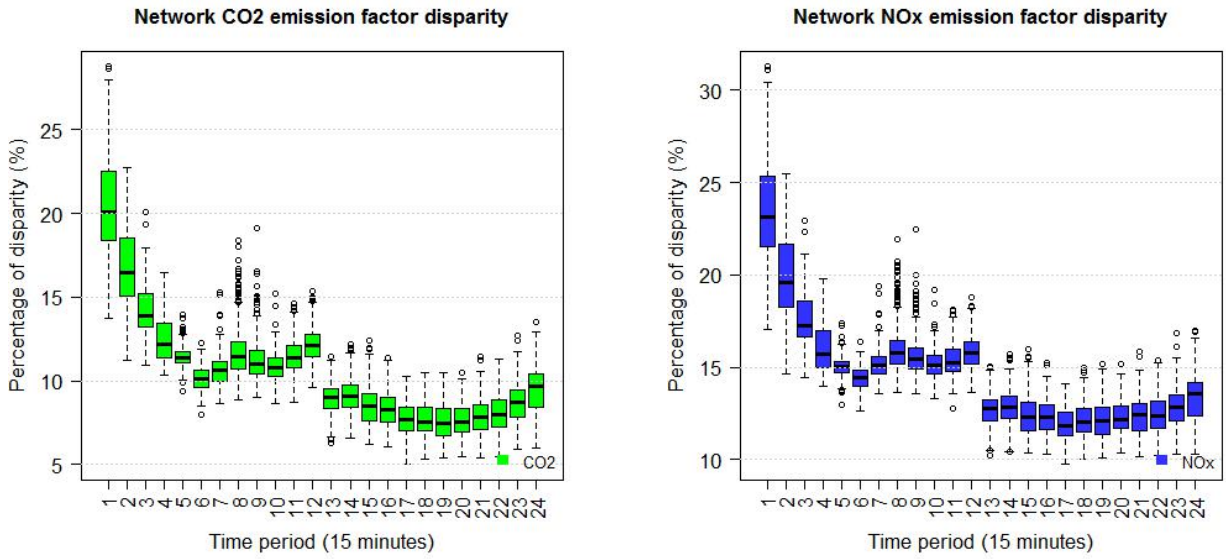
(b) $NO_x$ emission factor disparities by time period.

Fig. 5.8 – *Disparities on emission factors by time period.*

The emission factors are calculated for each pollutant by time period considering both mean speeds. The definition of the emission function curves leads to emission factors that are higher than those of the arithmetic speeds since the harmonic mean speeds are lower than the arithmetic definition. Hence the differences between both of them are always positive and vary between 4% and 26% on average for $CO_2$ and 3% and 20% on average for $NO_x$. Considering the traffic states, the gap increases in the congested periods and decreases in the free flow ones. The same happens with the data dispersion. Therefore, by taking into account only the first term of 5.17, a gap between both emission factors can be identified. They are derived from how the mean speed is considered in space and time, and the aggregated the data the more the mean speed values tend to be reduced. This leads to high coefficients for the COPERT emission functions and thus higher emission estimations.

To preserve the convexity of the function, a non-negative weighted sum is necessary and is represented by the second term. The second term increases the gap calculated in the first term of 5.17. Figure 5.9 shows the difference in percentage between $f(\tilde{V})$ and $\frac{\sum_i f(\bar{v}_i)}{n}$ by time period. The difference between these functions depends on the local mean speeds. Before running the simulations, the network is empty, with no car inside. Car demand starts in the first period and so the mean speeds are higher, thus the disparity in these periods is also larger and varies due to the fact that the car demand values are completely different in each simulation. This variation tends to stabilize through time because the volume of cars defined inside the network is reached and stays constant over the periods. Considering the morning assignment, i.e. periods from 1 to 12, the gaps for $CO_2$ are between 8% and 15% on average and for $NO_x$ they are between 13% and 18%. In the afternoon assignment,

the disparity is smaller and can vary between 5% and 13% for $CO_2$ and between 9% and 16% for $NO_x$.




(a) $CO_2$ emission factor disparities by time period.

(b) $NO_x$ emission factor disparities by time period.

Fig. 5.9 – *Percentage of disparity considering the convexity of the function.*

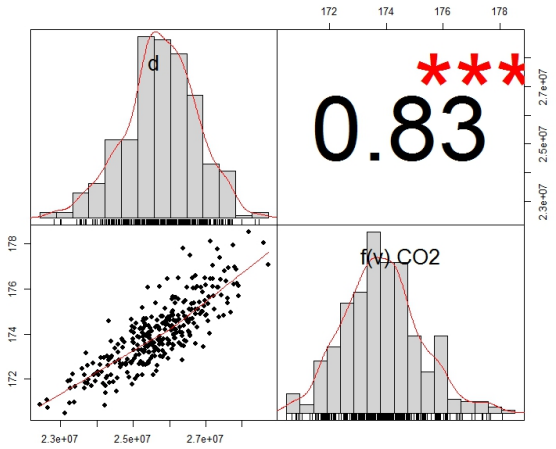
It can also be seen that the percentage of disparity of the $NO_x$ function is higher than that of $CO_2$. This is due to the fact that the degree of convexity of the $NO_x$ emission function is higher than that of the $CO_2$ function. The latter can be observed in figure 3. Finally, depending on the local speeds and the convexity of the pollutant emission functions, the second term of 5.17 can increase the disparity of emission estimations by 13% on average.

The last term of 5.17 reduces the disparity of the emission estimations. It depends on the number of links in the network and the covariance between the variables that describe the links, such as the traveled distances and emission factors. The covariance is positive for all the periods, so they move together in the same direction. To measure the degree to which the variables tend to move together, the correlations are calculated for each pollutant and are shown in figure 5.10.

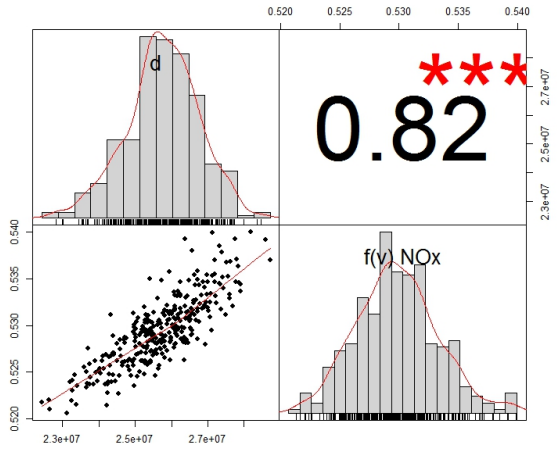
Figure 5.10 shows the distribution and the kernel density overlays on the diagonal, the bivariate scatter plot with a fitted line on the bottom of the diagonal, and finally the value of the correlation plus the significance level as stars. As can be seen, both variables are strongly correlated and show an uphill linear pattern. By considering the covariance as positive and correlated, it reduces the gap produced by the different spatial-temporal scales of the same traffic data.

As discussed in this section, the emission estimations using the COPERT emission functions can differ according to the spatial and temporal scales. The global scale considers the traffic data aggregated in space and time before calculating emissions, while calculating emissions in each link. The aim is to estimate the daily pollutant emissions of the network using both types of data. The emission functions are based on driving cycles and can precisely determine the emission factors for these scales. As COPERT emission functions are based on driving cycles, we conclude that the scale that provides the closest results correspond to relating is the local one as link information are comparable with driving cycles.

Evaluating the fact that emission functions are directly based on instantaneous speeds, the non-




(a) Correlation between traveled distance and emission factor of $CO_2$.




(b) Correlation between traveled distance and emission factor of $NO_x$.

Fig. 5.10 – *Correlation between traveled distance and emission factor.*

linear behavior of this variable and the convexity of the emission function, the finer the data described in space and time the closer the emission estimation at this level will be to the driving cycles. This means that the local scale is the best choice for evaluating the emissions. In our case study, the global approach tends to overestimate emissions by about 14% in general compared to the local scale. Regarding pollutant emissions, when considering carbon dioxide (i.e. $CO_2$ emissions), the overestimation can be between 10% and 18% while for nitrogen oxides (i.e. $NO_x$ emissions) the overestimation is between 11% and 16%. When considering traffic conditions, in free-flow conditions the disparities are around 4% for $CO_2$ and 7% for $NO_x$ emissions. In a congested situation, the overestimation can reach about 23% compared to the local approach.

## 5.3 Conclusion of the chapter

This chapter highlighted how spatial-temporal scales of the same traffic data can induce a bias on emission estimations based on emission-factor curves. In order to assess the latter, a microscopic traffic simulator was used to provide the finest traffic data of the 6$^{th}$ district of Paris to conduct the study. Two scales were defined: the first called local that described the traffic data at link level; and the second that considered the same traffic data at network level. Defining the emissions showed that there gaps existed between both scales that for the most part depended on the difference between average speeds, the convexity of the functions and the covariance between variables. These gaps can lead to overestimations of emissions of around 14% and as high as 23% for congested states.

The quality of traffic emission data input is obviously an important factor for the accuracy of emission estimations. In addition, the impact of the accuracy of traffic data on emission estimations using established emission factors appears to be an area that requires further work.

Considering the emission estimations based on emission factor curves, these curves already present uncertainty because they are based on the average values of several driving cycles. (Jaikumar et al., 2017) made a comparison between the real-world emission measurement system and the emission factors from COPERT, and the results showed that the real-world emissions were several times higher when compared to the emissions derived from the established factor curves. A part of these differences

can come from aggregation bias as revealed in this chapter but other factors can be involved.

# Conclusions

The objective of this thesis was to define methods for sampling data to estimate total emissions at several spatial and temporal scales from a sample. The motivation for this study is to improve the accuracy of the input data needed in emission models so that they can quantify road emissions reliably. This accuracy is related to traffic dynamics, since the volumes of pollutant emission are sensitive to several factors such as driver behavior (*i.e.* aggressiveness), traffic conditions (*i.e.* free-flow or congested), the operating mode of the vehicle (accelerations, decelerations and stop-and-go), the vehicle fleet, the type of road infrastructure, etc. In reality, this information can come from surveys and observations of field studies or from road technologies such as on-road sensors and probe vehicles. These methodologies are extremely costly and can represent the network only at the local level. Microscopic traffic simulations are used increasingly to better evaluate the transport networks. They allow assessing, making conclusions and testing new techniques without the need for disrupting real systems and performing new data collections. They also provide traffic information described in time and space, which is the great advantage of simulation models and why their use is so important.

However, the microscopic simulation of network traffics can be very time consuming and the volume of information that must be processed is not always computationally feasible which is precisely the point on which this thesis focused. Sampling was proposed to reduce the volume of data to be processed while maintaining accuracy and reliability. Using existing statistical methods instead of estimating emissions for each link in the network, the idea is to identify the most representative links of the network to estimate overall emissions. Identifying the right statistical methods that take into account spatial and temporal correlations of traffic data is important for the representativeness of the sample. The state of the traffic and its dynamics evolves in time and space and it is linked to changes of demands in the network as well as to peak periods. To this end, it is necessary to obtain good knowledge of the coupling between traffic and emission models.

The bibliography of this study identified the different levels of traffic models and their characteristics, and emission models and the relationships between them with different levels of detail relating to the variables. Existing models have been increasingly improved to better represent traffic behavior and to estimate the volumes of pollutants emitted into the atmosphere with accuracy and reliability.

The adequacy of the application of each type of model is fundamentally linked to the analysis of the objectives pursued. In general, studies that cover large areas and strategic analyses are considered more appropriate for use with macroscopic modeling tools, the overall results of which correspond to the objectives of the evaluation. On the other hand, microscopic modeling is best used in smaller

areas where detailed simulations are important for achieving objectives.

Regarding the analysis of emissions in large urban areas with heterogeneous infrastructures and operational characteristics, global models are unable to adequately represent the particularities of the traffic responsible for the largest changes in emissions, such as acceleration and deceleration events. On the other hand, microscopic models that have this capability, require large amounts of data that have to be correctly calibrated to present reliable estimates. Thus, the coupling of the microscopic traffic model and the static emission model proposed in this study proved to be adequate, as it retains the individual advantages of each approach and increases the quality of the results.

Regarding emission models, static models are suitable for strategic studies, such as the preparation of emission inventories in the main areas where average conditions are satisfactory. The COPERT methodology was identified as facilitating the applicability of coupling for the objectives proposed in the thesis and also for considering their application to real cases. This model is based on measurements of emissions during representative cycles and covers all the atmospheric pollutants of the various categories of vehicles represented by emission factors.

After defining the models to be used, a sensitivity analysis was carried out in order to evaluate the sensitivity of the model results when changing only one or more variables simultaneously and how this is reflected in the emission calculations. The most suitable data for such calculations were identified in the sampling methodology. The main results by chapter are described in the next section.

## 5.4 The results obtained

Chapter 2 provided a descriptive analysis of the traffic data, such as the distance traveled by vehicles and average speed, which were used as the input data in the COPERT IV emission model. The microscopic traffic simulator used allowed making comparisons of traffic data according to the source of information: punctual sensors that collect the type of traffic data used by city managers and spatial data from the microsimulation that gave a fine description of the vehicles inside the network. The study showed that the definition of the variables given by the punctual sensors overestimated the total distance traveled by vehicles by an average of 3%, and this bias could be further accentuated during periods of congestion, while average speed was overestimated by more than 100%. This bias was accentuated by the variability of the speeds of traffic in fluid state (*i.e.* free-flow). Considering the two definitions of traffic variables, emissions were calculated to observe how these biases influenced the quantification of emissions. Traffic data from the punctual sensors underestimate pollutant emissions by 14% on average. These factors highlight the considerable impact of simplifying traffic data used to calculate emissions. The spatial data from the microscopic simulator were used throughout the thesis to conserve the accuracy and reliability of the traffic data. Regarding the spatial data, this chapter also presented the study of correlations, the identification of links and the partition of the network according to certain characteristics of the variables supporting the sampling study.

Chapter 3 dealt with the first sampling method used, called LASSO. In this chapter the datasets used throughout the study were defined to represent several spatial and temporal scales. The method was able to deal with highly variable data using by taking advantage of the strong correlations between the variables and using a small set of available observations by taking into account the number of available predictors. Many possible sample sizes were proposed by LASSO for all the traffic and emission variables. They were compared, combined, and evaluated, by considering estimation errors and model reliability. Depending on the temporal scale, the sampling rate could vary from 5% to

30% of the network links. Among all possibilities proposed in the chapter, the method was able to select links and use them to build a model to estimate the variables at network level with an average percentage error between 5% and 10%. This sampling method proved efficient for data processing in terms of computation effort and time.

Chapter 4 presented other statistical methods used to perform sampling selection for emission data only in two different temporal ranges using the same data as that of the previous chapter. Two approaches were proposed: (i) the use of naive sampling methods to estimate emissions under the same conditions and rate selection as LASSO did in the previous chapter. They were applied on two spatial scales: the entire network and zones defined by the partitioning methodology, and (ii) a sampling method that selected its own group of links based on their statistical significance in the linear regression model. The aim was to test the quality of the results, the computation time and the influence of the errors in the emission estimations on different spatial and temporal scales. A comparison between the model results was made and showed that clustering the network improved the quality of the emission estimation in comparison to the other methods. Despite considering a lower selection rate compared to LASSO, the emission estimation error did not exceed ±5%. A real case of application was considered in the end of the chapter which used the simulated data only from the links currently equipping the $6^{\text{th}}$ district of Paris. 20% of the network is equipped by punctual sensors, so the daily values of emissions can be estimated with less than ±2% error, and considering a shorter temporal scale for real-time tracking, the emission were estimated with less than ±10% error.

Chapter 5 presented a study of the impact of spatial and temporal aggregation of traffic data on the emission estimation. Biases were observed when comparing the emissions estimated from links (*i.e.* local scale) and from network (*i.e.* global scale) using the same traffic data. As a result, the more the traffic data were aggregated in space and time, the greater the bias. In addition, these disparities were strongly influenced by traffic states. The overestimation for congested states was 23% on average. The origin of this bias was quantified as a function of the traffic indicators and demonstrated it stemmed from the emission, convexity functions and the covariance of the traffic variables.

Through these various studies, this work allowed me to identify the stakes associated with the estimating emissions on the urban level. Among the studies proposed, the importance and influence of aggregated and disaggregated traffic approaches, with respect to coupling traffic and emission models, proved to be an essential point for ensuring the reliability of the results. The main precautions to be taken into account when coupling are the consistency of road network representations, the flows, and the dynamic traffic conditions. The calibration of the models was also of great importance for ensuring the quality of the results. Furthermore, it is important to underline that static emission models such as COPERT, which estimates emissions based on emission-factors, can also introduce a misrepresentation of the emissions when considering the scale of application. This is due to the difference between the standardized driving cycles integrated in the model and the real driving cycles, which lead to biased emission levels.

Regarding the operational dimension, the challenge, therefore, is to provide improved air quality without compromising the mobility of the population. Confronted by the problem of missing data and the difficulty of obtaining a complete representation of the traffic of an entire transport network, the sampling methodology can help to assist in choosing the most effective traffic strategy to be adopted for each type of problem being treated. It can also assess the performance of traffic management strategies to reduce pollutant emissions. The applications of such techniques are numerous. In addition to significant improvement in computing time, the development of appropriate sampling methods could

also help to identify key areas of a network, improve assessments a posteriori (optimal positioning of measurement instruments such as on-road sensors, the definition of reference vehicles trips with on-board measurements). The methodologies covered by this work could also be useful in real-time emission quantification assessments.

It is important to emphasize that the objective of this work was not to evaluate the absolute values of the emissions, but to identify the influence and variability in the estimates due to differences in the application of the models and the input data. However, it is important to note that the emission factors used by the emission models in this thesis should be based on vehicle measurements from the study area. However, it is recognized that obtaining accurate measurements of emission factors represent a complex and costly process.

The relevance of coupling traffic and emission models allied with the sampling method can also be enhanced by the fact that they are an important aid in the decision-making process, as they can provide large quantities of data in a systematic and reproducible way, allowing the a priori comparison of several alternative scenarios without it being necessary to implement them. In this way, the methodology provides significant operational advantages when applied as an evaluation tool, although account should also be taken of its disadvantages. Among the main advantages are the following:

- it provides a means for addressing complex problems arising from the many existing interrelationships and the sensitivity of road emission to the variability of a road system;
- it gives greater freedom to experiment/test/evaluate traffic and emission strategies;
- it allows faster, more flexible and less costly analyses.

Regarding its operational disadvantages, the following points should be noted:

- whatever the traffic model and emission model used, however complex they may be, they essentially simplify real conditions;
- the methodology involves a large number of variables that have to be estimated;
- the application of the methodology requires significant resources and materials, and people with specialized technical skills, so that it can be used and evaluated with the necessary rigor.

The method developed in this thesis presents great potential for formulating emission diagnostics in urban areas and for the evaluating different road performance scenarios. Finally, the method proposed in this work may spur other studies on this topic, as specific studies will lead to results that can contribute to its improvement.

## 5.5 Perspectives

As a recommendation for future work, the application of the proposed method is proposed in different contexts, involving urban areas with different road infrastructures and operating characteristics and network sizes and heterogeneities. Moreover, although this research was limited to studying only passenger cars and hot exhaust emissions, different vehicle fleet compositions and types of fuels can be incorporated into the method. Also, coupling the microscopic traffic model with other emission models would be profitable and be a way of ascertaining the efficiency and effectiveness of the methodology for different networks, and verifying if it can be used in this way.

Indeed, the sensor networks used to measure air pollution are of generally low density and do not specifically discriminate the contribution of road traffic emission. Conversely, real-time traffic data and especially probe vehicles are developing enormously. An interesting idea would therefore be to supplement the measurements by dedicated sensors with information from probe vehicles coupled with an emission model. Once again, identifying relevant traffic information through appropriate sampling methods makes sense.

# References


Ahn, K. (2002). Modeling light vehicle emissions based on instantaneous speed and acceleration levels. Doctoral thesis, Virginia Polytechnic Institute and State University.

Andre, M., Hickman, A., Hassel, D., and Journard, R. (1995). Driving cycles for emissions measurements under european conditions. Technical paper, SAE publications.

Ang, A. and Tang, W. (2007). *Probability concepts in engineering*. John Wiley and Sons Inc.

Barcelo, J., Ferrer, J. L., Garcia, D., Grau, R. Forian, M., and Chabini, I. Le Saux, E. (1998). Microscopic traffic simulation for att systems analysis: a parallel computing version. Contribution to the 25th aniversary of crt.

Barlow, T. J., Latham, S., McCrae, I. S., and Boulter, P. G. (2009). *A reference book of driving cycles for use in the measurement of road vehicle emissions*. TRL Limited.

Barth, M., Na, F. Younglove, T., Scora, G., Levine, C., Ross, M., and Wenzel, T. (2000). Development of a comprehensive modal emission model. Final report, National Research Council.

Bien, J. (2016). Model selection using the lasso technique. Online URL: https://www.cscu.cornell.edu/workshops/lasso.php, accessed July 12, 2016.

Boxil, S. A. and Yu, L. (2000). An evaluation of traffic simulation models for supporting its development. Report, Center for transportation training and research, Texas Southern University.

Burghout, W. (2004). Hybrid microscopic-mesoscopic traffic simulation. Doctoral thesis, Royal Institute of Technology, Stockholm.

Can, A. and Leclercq, L. (2009). Estimation des consommations énergétiques et des polluants émis par le trafic routier : Revue bibliographique des modèles existants. Technical report, IFSTTAR.

Cappiello, A. (2002). Modelling traffic flow emissions. Master of science in transportation, Massachussetts institute of technology.

CARB (2002). Emfac: California air resources board's emission inventory serie. Online URL:www.arb.ca.gov, Accessed June 01, 2017.

Carslaw, D. C. and Beevers, S. D. (2005). Estimation of road vehicle primary no2 exhaust emission fractions using monitoring data in london. *Atmospheric Environmental*, 39:167–177.

Cassotti, M. and Grisoni, F. (2011). Variable selection methods: an introduction. Tutorial report 6, University of Milano - Bicocca.

Chakroborty, P. (2006). Models of vehicular traffic: An engineering perpective. *Physica*, 372:151–161.

Chatterjee, A., Miller, T., Philpot, J., Wholley, T., Guensler, R., Hartgen, D., Margiotta, R., and Stopher, P. (1997). Improving transportation data from mobile source emission estimates. In *NCHRP report No. 394 on Transportation Research Boarder*.

Daganzo, C. F. (1980). Optimal sampling strategies for statistical models with discrete dependent variables. *Transportation Science*, 14:324–345.

Daganzo, C. F. (1994). The cell transmission model : A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B*, 28:269–287.

De Nevers, N. (2000). *Air pollution control engineering*. New York: McGraw-Hill, 2 edition.

De Vito, E. and Veronica Umanità, S. V. (2011). A consistent algorithm to solve lasso, elatic-net and tikhonov regularization. *Journal of Complexity*, 27(2):188–200.

Department of Statistics Online Programs, The Pennsylvania State University (2016). Stepwise regression. Online URL: https://onlinecourses.science.psu.edu/stat501/node/329, accessed December 10, 2016.

Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Styatistical Psychology*, 45:265–282.

DieselNet (2017). FTP-75. https://www.dieselnet.com/standards/cycles/ftp75.php. [Online; accessed 05-April-2017].

DOE (2015). Australia's emissions projections. Report 2014-15, departament of the Environment.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley-Interscience.

EEA (2016). Use of cleaner and alternative fuels. Online URL: https://www.eea.europa.eu/data-and-maps/indicators/use-of-cleaner-and-alternative-fuels/use-of-cleaner-and-alternative-10, Accessed July 10, 2017.

EEA (2017a). Size of the vehicle fleet in eu. Online URL: https://www.eea.europa.eu/data-and-maps/indicators/size-of-the-vehicle-fleet/size-of-the-vehicle-fleet-7, Accessed July 09, 2017.

EEA (2017b). Transport emissions of greenhouse gases in eu. Online URL: https://www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-10, Accessed July 09, 2017.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Efroymson, M. A. (1960). *Matematical Methods for Digital Computers*. Wiley.

Elloumi, N., Haj-Salem, H., and Papageorgiou, M. (1994). Metacor: A macroscopic modelling tool for urban corridor. In triennal symposium on transport analysis.

EPA (2004). Mobile source emission factor model. Epa report epa-420-r-03-010, US Environmental Protection Agency.

EPA (2010). Motor vehicle emission simulator: Moves 2010 user guide. Epa report epa-420-b-09-041, US Environmental Protection Agency.

EU (2013). Environmental action programme to 2020. Online URL: http://ec.europa.eu/environment/action-programme/, Accessed July 09, 2017.

Fellendorf, M. and Vortisch, P. (1998). *Microscopic Traffic Flow Simulator VISSIM*, volume 145. Springer.

Fellendorf, M. and Vortisch, P. (2000). Integrated modeling of transport demand, route choice, traffic flow and traffic emissions. Conference paper, 79th Annual Meeting of the Transportation Research Board.

Flomn, P. L. and Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should to use. In *NESUG*.

Font, A. and Fuler, G. W. (2016). Did policies to abate atmospheric emissions from traffic have a positive effect in london? *Environmental Pollution*, 218:463–474.

Frecht, D., Fischer, P., Fortunato, L., Hoek, G., De Hoogh, K., Marra, M., Kruize, H., Vienneau, D., Beelen, R., and Hansell, A. (2015). Associations between air pollution and socioeconomic characteristics, ethnicy and age profile of neighborhoods in england and the netherlands. *Environmental Pollution*, 198:201–210.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Geroliminis, N. and Daganzo, C. F. (2007). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Working paper, UC Berkeley Center for Future urban Transport.

Ghosh, S. (2007). Adaptative elastic net: An improvement of elastic net to achieve oracle properties. Technical Report PR-07-01, Indiana University.

Gkatzoflias, D., Kouridis, C., Mellios, G., and Ntziachristos, L. (2012). Copert 4: Computer programme to calculate emissions from road transports. Technical report, Laboratory of Applied Thermodinamics Mechanical Engineering Department, University of Thessaloniki.

GLA (2010). Clearing the air: the mayor's air quality strategy. Report 176, Greater London Authority.

Hastie, T. and Efron, B. (2009). Lars: Least angle regression, lasso and forward stagewise. R package version 0.9-7, Stanford University.

Hastie, T., Tibshirani, R., and Friedman, J. (2009a). *The Elements of Statistical Learning*. Wiley series in survey methodology. Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009b). Glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.1-4,, Stanford University.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The elements of Statistical Learning: Data mining, Inference and Prediction*. Springer.

Hausberger, S., Reixes, M., Zallinger, M., and Luz, R. (2009). Emission factor from the model phem for the hbefa version 3. Technical report no i-20a/2009, Graz University of Technology, Institute for Internal Combustion Engines and Thermodynamic.

Haworth, J. and Cheng, T. (2011). Graphical lasso for local spatio-temporal neighbourhood selection. Technical report, University College of London.

Hoogendoorn, S. P. and Bovy, P. H. L. (2001). State-of-the-art of vehicular traffic flow modelling. *Journal of Systems and Control Engineering*, 2015:283–303.

Huang, J., Ma, S., and Zhang, C. (2006). Adaptive lasso for sparse high-dimensional regression models. Technical Report 374, Department of Statistics and Actuarial Science, University of Iowa.

Ichikawa, S. M., Pitombo, C. S., and Kawamoto, E. (2002). Application of data mining in obtaining relationship between chained travel patterns and socio-economic characteristics. In *XVI ANPET Annual Meeting*.

IPCC (2013). Climate change 2013: The physical science basis. The fifth assessment report of the intergovernmental panel on climate change, Cambridge University.

ITS (1993). Dracula: Dynamic route assignment combining user learning and microsimulation. Online URL:www.its.leeds.ac.uk/software/ dracula/, Accessed June 02, 2017.

Jaikumar, R., Shiva Nagendra, S. M., and Sivanandan, R. (2017). Modeling of real time exhaust emissions of passengers cars under heterogeneous traffic conditions. *Atmospheric Pollution Research*, 8:80–88.

Ji, Y. and Geroliminis, N. (2012a). On the spatial partitioning of urban transportation network. *Transportation Research Part B*, 46:1639–1656.

Ji, Y. and Geroliminis, N. (2012b). On the spatial partitioning of urban transportation networks. *Transportation Research Part B*, 46:1639–1656.

Jiang, H. (2009). A matlab implementation of glmnet. Technical report, Stanford University.

Jie, L., Zuylen, H. V., Chen, Y., Viti, F., and Wilmink, I. (2013). Calibration of a microscopic simulation model for emission calculation. *Transportation Research Part C*, 31:172–184.

Kamarianakis, Y., Shen, W., and Wynter, L. (2012). Real-time road traffic forescasting using regime-switching space-time models and adaptive lasso. *Applied Sthochastic Models in Business and Industry*, 28:297–315.

Katsis, P., Ntziachristos, L., and Mellios, G. (2012). Description of new elements in copert 4 v10.0. Technical report, EMISIA SA Report.

Koenker, R. and Mizera, I. (2014). Convex optimization in r. *Journal of Statistical Software*, 60(5).

Kouridis, C., Gkatzoflias, D., Kioutsioukis, J., Ntziachristos, L.and Pastorello, C., and Dilara, P. (2010). *Uncertainty estimates and guidance for road transport emission calculations*. European Union.

Leblanc, D. C., Saunders, F., Meyer, M. D., and Guensler, R. (2005). Driving pattern variability and impacts on vehicle carbon monoxide emission. *Transportation Research Record*, 1472:45–52.

Leclercq, L., Laval, J., and Chevallier, E. (2007). The lagrangian coordinate system and what it means for first order traffic flow models. 17th international symposium on transportation and traffic theory (isttt), Elsevier Science.

Li, L., Su, X., Wang, Y., Lin, Y., Li, Z., and Li, Y. (2015). Robust casual dependence mining in big data network and its application to traffic flow predictions. *Transportation Reserach Part C*, 58:292–307.

Lighthill, M. J. and Whitham, G. (1955). On kinematic waves : Ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society*, 229:317–345.

Lopez, C., Krishnakumari, P., Leclercq, L., Chiabaut, N., and Van Lint, H. (2017). Spatio-temporal partitioning of transportation network using travel time data. In *TRB 2017 Annual meeting of Transportation Research Board*.

Mairie de Paris, D. d. l. V. e. d. D. (2015). Référentiel géographique pour les données trafic issues des capteurs permanents.

Mark, J. and Goldberg, M. A. (2001). Multiple regression analysis and mass assessment: A review of issues. *The Appraisal Journal*, Jan:89–109.

McTrans (2017). Corsim: Microscopic traffic simulation model. Online URL:http://mctrans.ce.ufl.edu/featured/tsis/version5/corsim50.htm, Accessed May 22, 2017.

Mukherjee, P. and Viswanathan, S. (2001). Carbon monoxide modeling from transportation sources. *Chemosphere*, 45:1071–1083.

Nau, R. (2017). Stepwise and all-possible-regressions. Online URL: https://people.duke.edu/ rnau/411home.htm, Accessed December 11, 2016.

Nicolas, R. (2013). The different driving cycles. http://www.car-engineer.com/the-different-driving-cycles/. [Online; accessed 05-April-2017].

Ntziachristos, L. and Samaras, Z. (2014). Emep/eea emission inventory guidebook. Guidebook, European Environmental Agency.

Ortuzar, J. and Willumsen, L. G. (2011). *Modelling transport*. John Wiley and Sons Ltd., 4 edition.

Pendse, G. V. (2011). A tutorial on the LASSO and the" shooting algorithm". *URL http://www. gautampendse. com/software/lasso/webpage/lasso_shooting. html*.

Pierson, W., Gertler, A., Robinson, N., Sagebiel, J., Zielinska, B., Bishop, G., Stedman, D., Zweidinger, R., and Ray, W. (1996). Real-world automotive emissions - summary of studies in the fort mchenry and tuscarora mountain tunnels. *Atmospheric Environment*, 30:2233–2256.

Rakha, H. and Ding, Y. (2003). Impact of stops on vehicle fuel consuption and emissions. *Journal of Transportation Engineering*, 129:23–32.

Richards, P. J. (1956). Shock waves on the highway. *Operations Research*, 4:42–51.

Rocha, T. V. (2013). Quantification des erreurs associées à l'usage des trajectoires simplifieés, issue de modèles de trafic, pour le calcul de la consommation en carburant. Doctoral thesis, L'Ecole Nationale de Travaux Publics de l'Etat.

Rouphail, N. M., Frey, J. D., Colyar, A., and Unal, A. (2000). Vehicle emissions and traffic measures: exploratory analysis of field observations at signalized arterials. Conference paper, 79th Annual Meeting of the Transportation Research Board.

Saeedmanesh, M. and Geroliminis, N. (2015). Clustering of heterogeneous networks with directional flows based on "snake" similarities. *Transportation Research Board*.

Scherr, W. (2003). An integrated model for planning and traffic engineering. Conference paper, Transportation Research Board.

Schiper, N., Lejri, D., and Leclercq, L. (2016). How selection techniques on traffic data sets can help in estimating network vehicle emissions. *Journal of Earth Sciences and Geotechnical Engineering*, 6:51–69.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2).

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–9052.

Simanaitis, D. J. (1977). Emission test cycles around the world. *Automotive Engineering*, 85:34–43.

Smit, R., Ntziachristos, P., and Boulter, P. (2010). Validation of road vehicle and traffic emission models - a review and meta-analysis. *Atmospheric Environment*, 44:2943–2953.

Smit, R., Poelman, M., and Schrijver, J. (2008). Improved road traffic emission inventories by adding mean speed distributions. *Atmospheric Environment*, 42:916–926.

Stanford, S. . F. C. (2016). Bayesian information criterion. Online URL: http://www.stanfordphd.com/BIC.html, accessed July 15, 2016.

Stathopoulos, A. and Karlaftis, M. (2001). Temporal and spatial variations of real-time traffic data in urban areas. In *80th TRB Annual Meeting*.

Sturm, P. J., Pucher, k., and Sudy, C. Almbauer, R. A. (1996). Determination of traffic emissions – intercomparison of different calculation methods. *The Science of Total Environment*, 189:187–196.

SYSTRA (2016). Paramics microsimulation. Online URL:http://www.sias.com/2013/sp/sparamicshome.htm, Accessed May 22, 2017.

Sétra (2012). Évaluation environnementale des projets de gestion dynamique de trafic - la qualité de l'air. Rapport technique, Service d'études sur les transports, les routes et leurs aménagements.

TfL (2014). Travel in london. Report 7, Transport for London.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288.

Tsanakas, N., Ekstrom, J., and Olstam, J. (2017). Reduction of errors when estimating emissions based on static traffic model outputs. *Transportation Research Procedia*, 22:440–449.

UNFCCC (2012). Kyoto protocol. Online URL: http://unfccc.int/kyotoprotocol/items/2830.php, Accessed July 10, 2017.

UNFCCC (2014). United nations framework convention on climate changes. Online URL: http://unfccc.int/2860.php, Accessed July 09, 2017.

Van Aerde, M. (1999). Integration release 2.20 for windows. User's guide, MVA and Associates.

Villegas, D., Canaud, M., and Bécarie, C. (2013). Constitution d'un environnement de simulation à grande échelle et production de données synthétiques pour le trafic routier et étude des données mobiles pour l'analyse des déplacements piétons- ISpace&Time. Technical report D4.4, IFSTTAR, Paris.

Vlieger, I., Keukeleere, D., and Kretzchmar, J. (2010). Environmental effects of driving behaviour and congestion related to passenger cars. *Atmospheric Environment*, 34:4649–4655.

Wenzel, T., Singer, B. C., and Slott, R. (2000). Some issues in the statistical analysis of vehicle emissions. *Journal of Transportations Statistics*, 3:1–14.

WHO (2013). Review of evidence on health aspects of air pollution. Technical Report REVIHAAP first results, World Health Organization.

Yu, L. (1998). Remote vehicle exhaust emission sensing for traffic simulation and optimization models. *Transportation Research Part D*, 3:337–347.

Zehe, D. (2015). Multi-resolution traffic modeling. Online URL:http://www.rp5.info/blog/2015, Accessed May 17, 2017.

Zhang, M. (2011). Reducing transportation emissions through land use intervention : Potentials and challenges. *Transportation Research Record*, 1729:18.

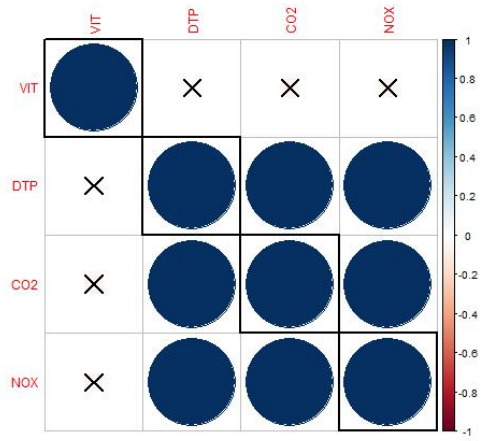
Zou, H. and Hastie, T. (2005). Regularization and variable selection via elastic net. *Journal of the Royal Statistical Society*, 67:301–320.

# Appendices

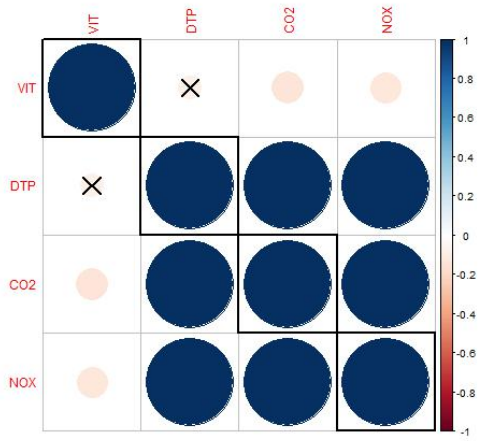
# A

## Correlations between variables by period of time

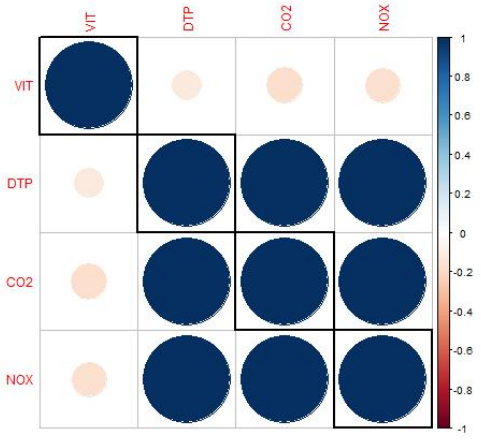
The correlation between variables through periods of time is presented below. For further analysis, go to the subsection Correlations between variables in 2.3.4.



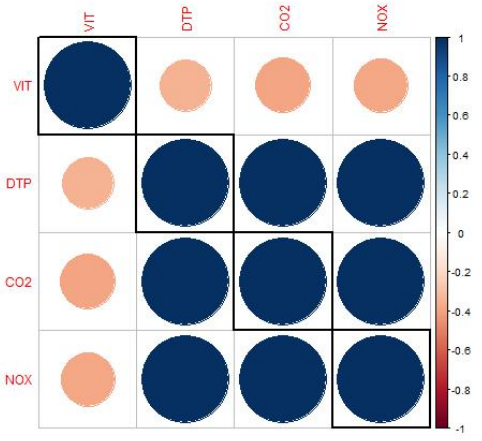

(a) Variable correlations in period 1



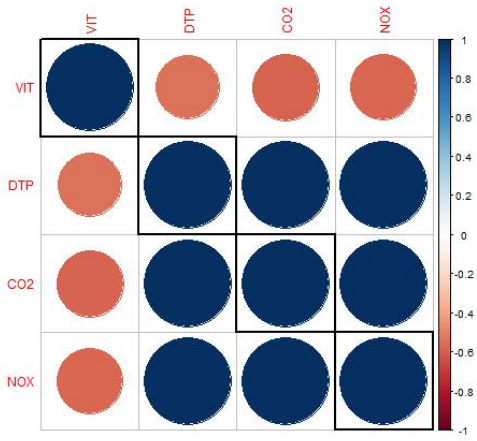

(b) Variable correlations in period 2



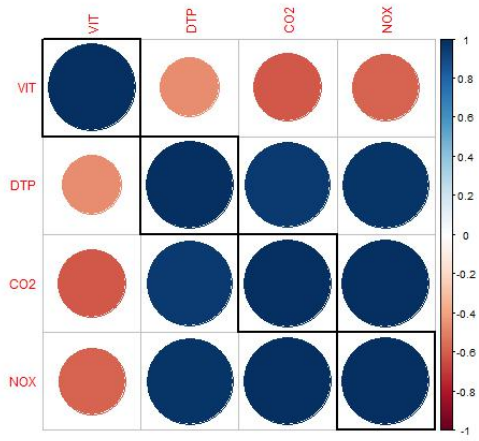

(c) Variable correlations in period 3



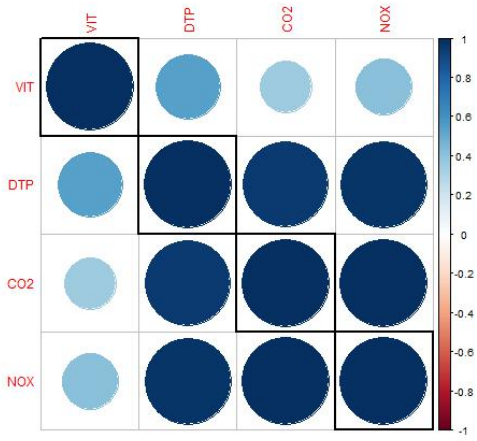

(d) Variable correlations in period 4



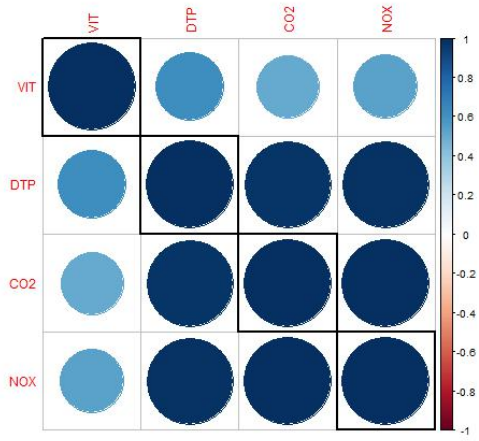

(e) Variable correlations in period 5



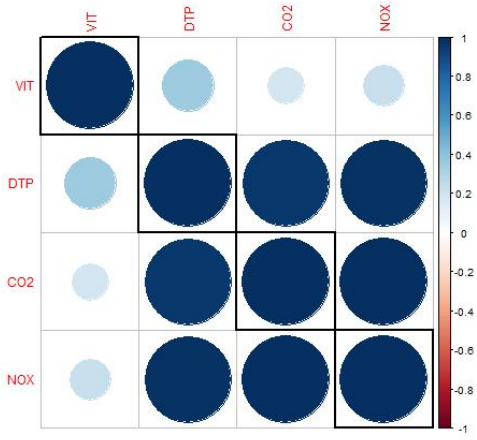

(f) Variable correlations in period 6



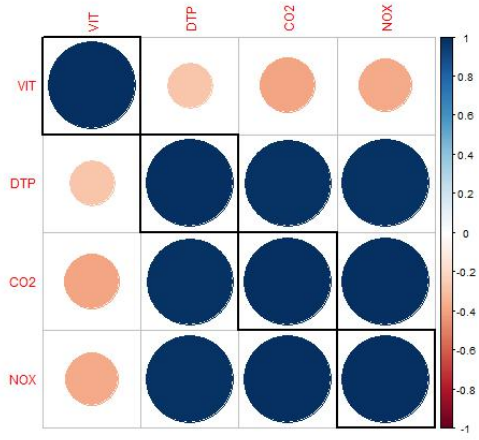

(g) Variable correlations in period 7



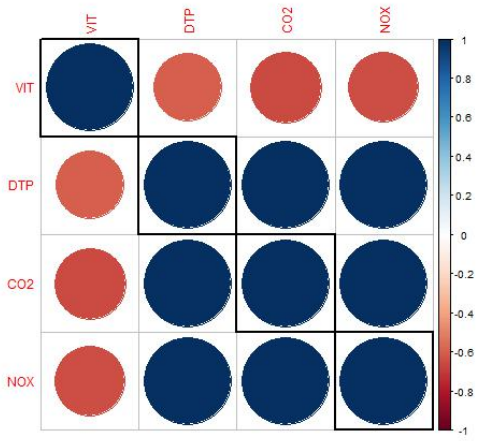

(h) Variable correlations in period 8



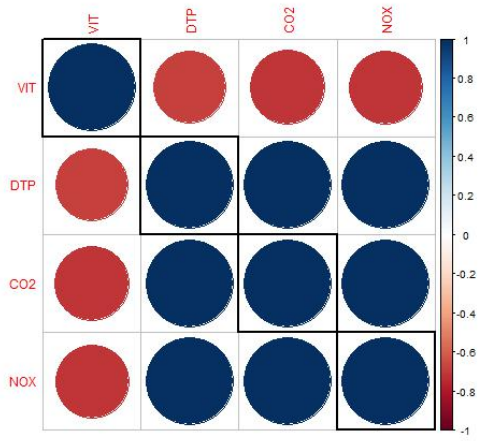

(i) Variable correlations in period 9



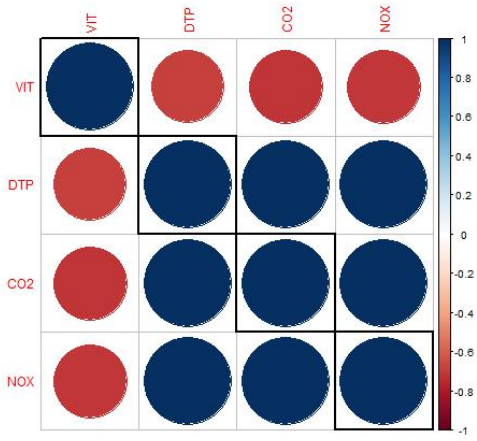

(j) Variable correlations in period 10



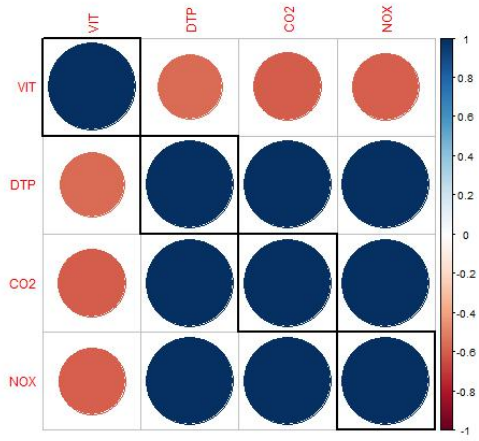

(k) Variable correlations in period 11



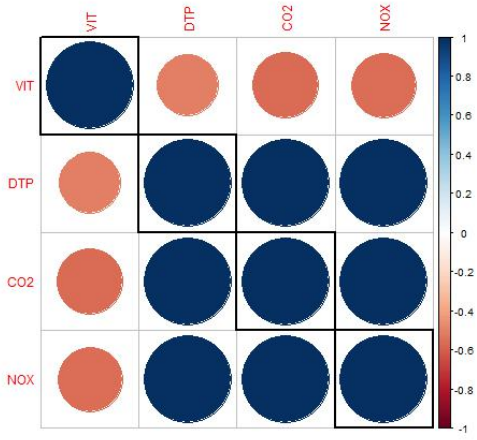

(l) Variable correlations in period 12



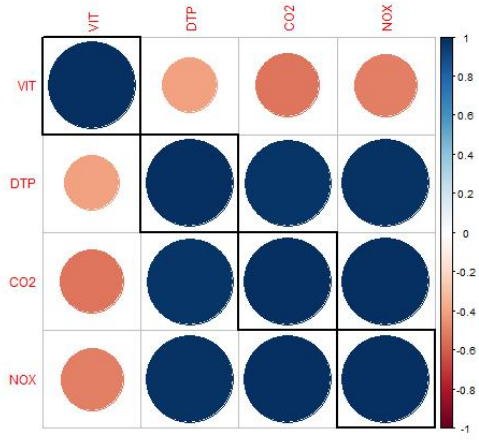

(m) Variable correlations in period 13



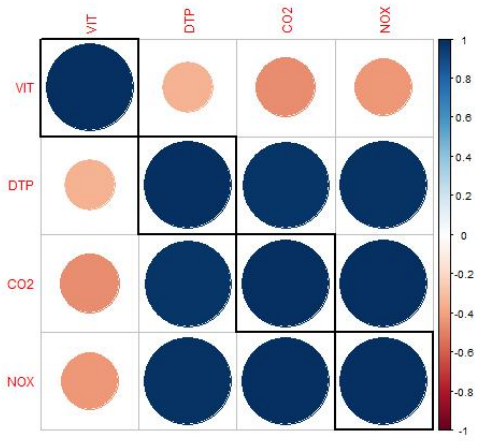

(n) Variable correlations in period 14



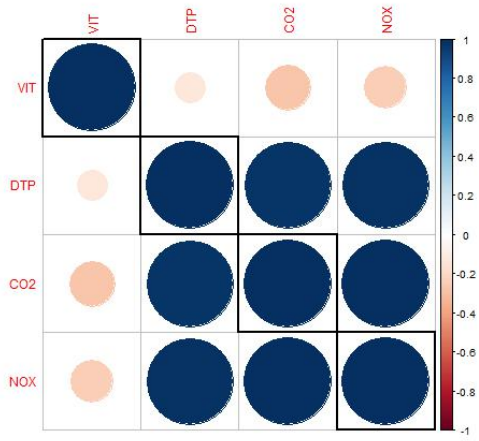

(o) Variable correlations in period 15



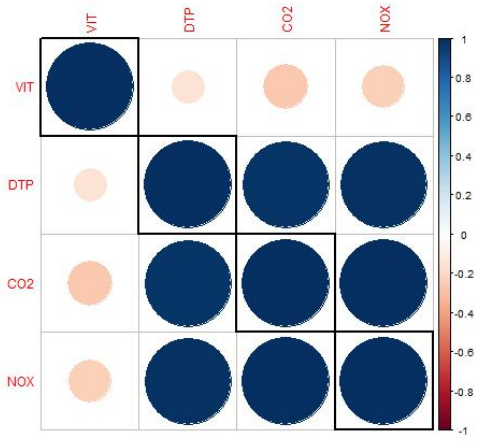

(p) Variable correlations in period 16



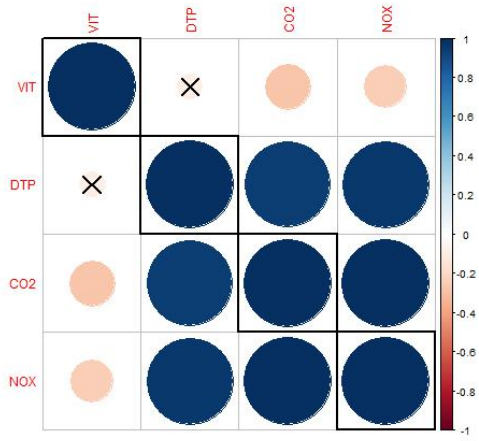

(q) Variable correlations in period 17



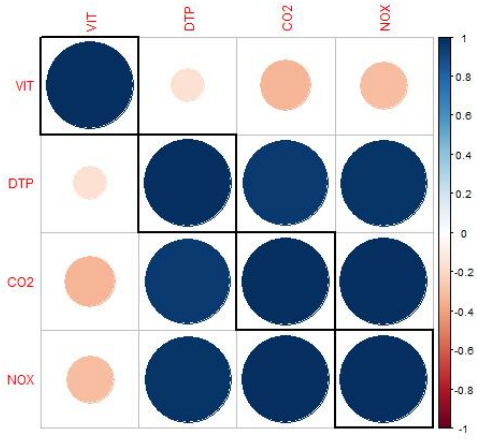

(r) Variable correlations in period 18



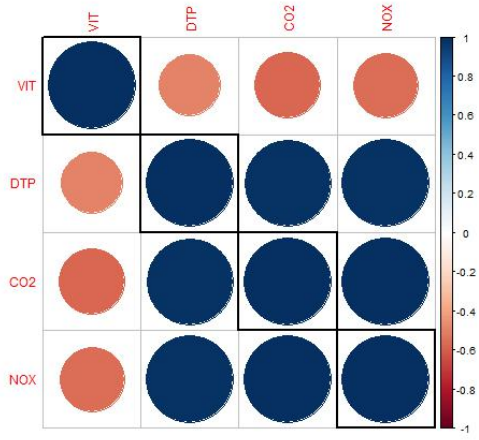

(s) Variable correlations in period 19



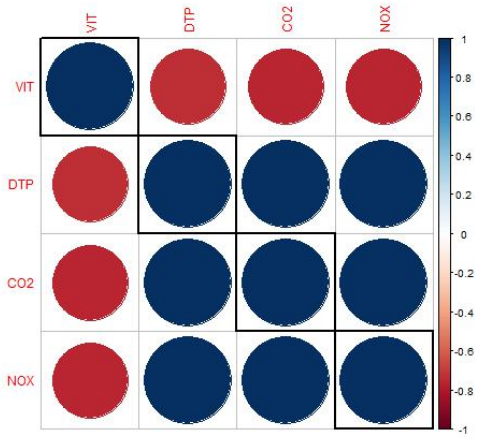

(t) Variable correlations in period 20



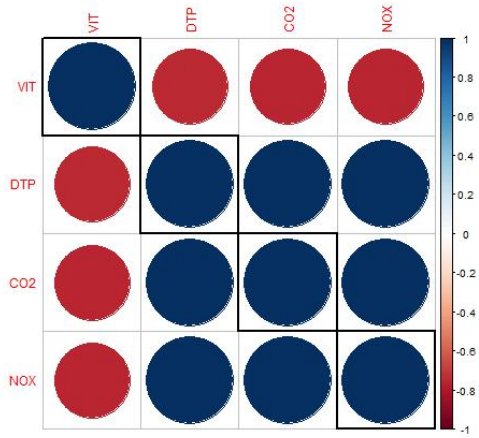

(u) Variable correlations in period 21




(v) Variable correlations in period 22




(w) Variable correlations in period 23




(x) Variable correlations in period 24

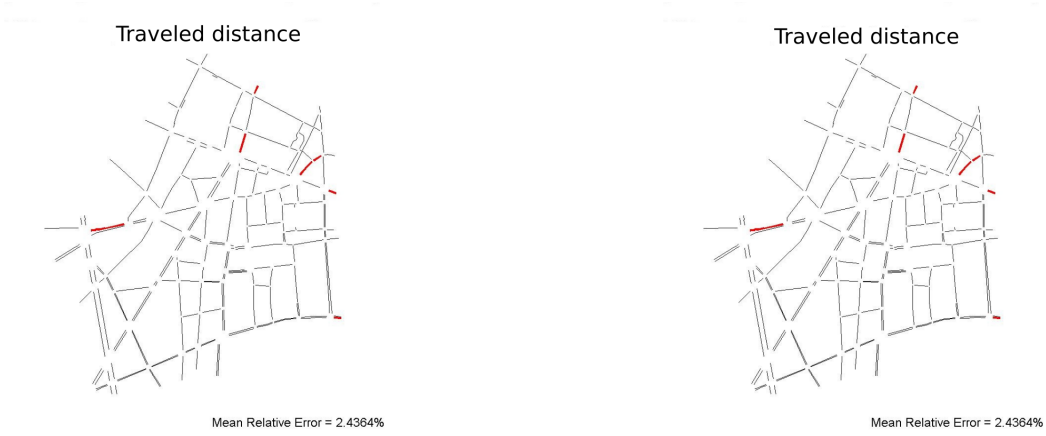
Fig. A.1 – *Variable correlations through periods of time*

# B

# Static/Static dataset appendix

## B.1 Stability of LASSO selection

Considering the static/static dataset and its variables, each link has 400 observation values. These observations were split up randomly into a training set and a validation set. The training set corresponded to 2/3 and the validation set to 1/3 of the observations. The LASSO method was applied to the training set and the selection made by it was validated in the validation set. To test the stability of the selection, two random sets were built for each variable to apply the shrinkage method and compare their results. The aim was to see if we had the same selection or if they were completely different. The same data was used for the two sets but organized differently. The following figures show the selection results for both random sets and for each variable considering the model determined by $\lambda$, with one standard error from the minimum.




(a) Selected links in traveled distance - random set 1.

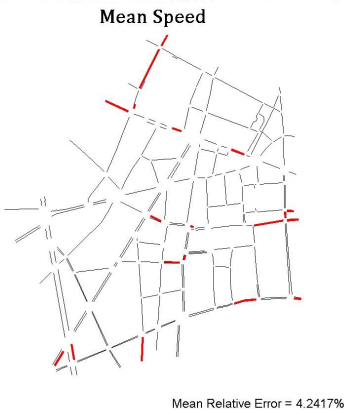
(b) Selected links in traveled distance - random set 2.

Fig. B.1 – *Comparison between selections made for different training and validation sets in travel production.*

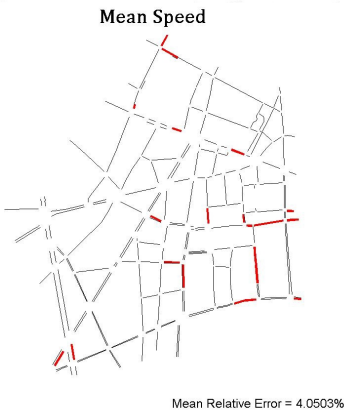
For travel production, both random sets had 7 selected links with 4 identical links present in both. For the mean speed variable, both random sets had 19 selected links of which 13 were the same. For the $CO_2$ emissions, both random sets had 11 selected links of which 7 were the same in both. For the $NO_x$ emissions, random set 1 had 16 selected links while set 2 had 11. 8 of the links were the same in both sets.

The method was run twice with the same dataset and the same results were obtained. Over half

of the selected links were the same in both random sets. The difference between their links was due to the fact that different observation values were used in the training sets but the correlations between them were the same in both sets.

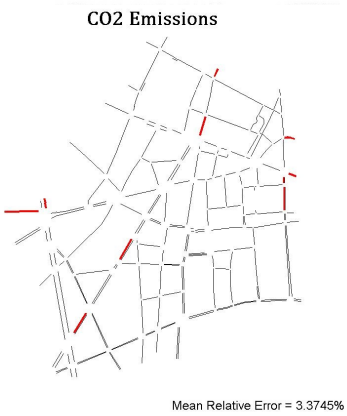



(a) Links selected for spatial mean speed - random set 1.

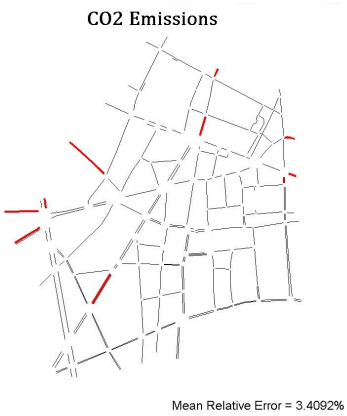



(b) Links selected for spatial mean speed - random set 2.

Fig. B.2 – *Comparison between selections made for different training and validation sets for mean speed.*




(a) Links selected for $CO_2$ emissions - random set 1.




(b) Links selected for $CO_2$ emissions - random set 2.

Fig. B.3 – *Comparison between selections made for different training and validation sets for $CO_2$ emissions.*




(a) Links selected for $NO_x$ emissions - random set 1.

(b) Links selected for $NO_x$ emissions - random set 2.

Fig. B.4 – *Comparison between selections made for different training and validation sets for $NO_x$ emissions.*

## B.2 Model size and the error distribution for static/static datasets

### B.2.1 Traveled distance

Figure B.5 shows the evolution of error distributions according to the lambda defined by LASSO for the traveled distance variable. All the conclusions were explained in 3.3.1.

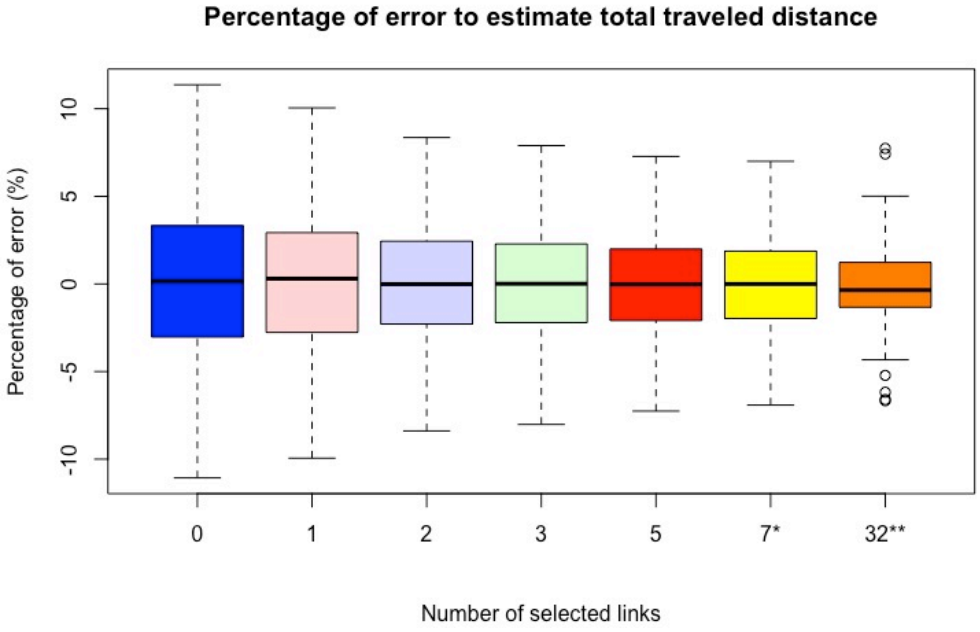



Fig. B.5 – *The percentage error of models calculated with different lambdas.*

Figure B.6 shows the same type of traveled distance model between 1SE lambda (*) and lambda (**).

Figure B.7 shows the error distribution of traveled distance of models that had more selected predictors than the lambda model.

### B.2.2 Spatial mean speed

Figure B.8 shows the error distributions from the model with fewer predictors than the 1SE lambda.

Figure B.9 presents the models with more predictors than the 1SE lambda model.

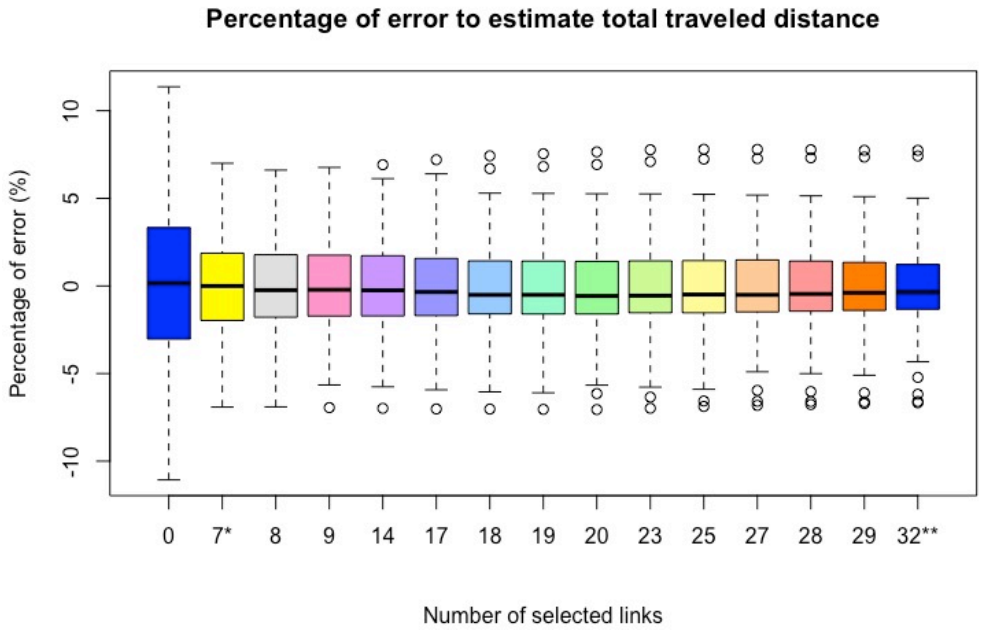



Fig. B.6 – *The percentage error of models between 1SE lambda and lambda.*

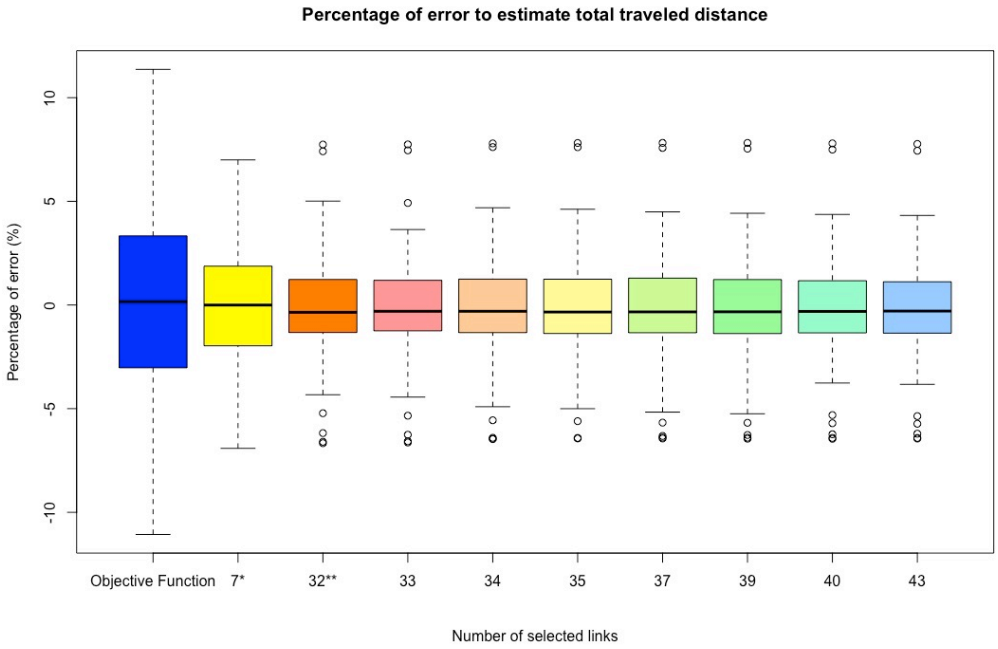



Fig. B.7 – *The percentage error from models with more predictors than the optimized models.*

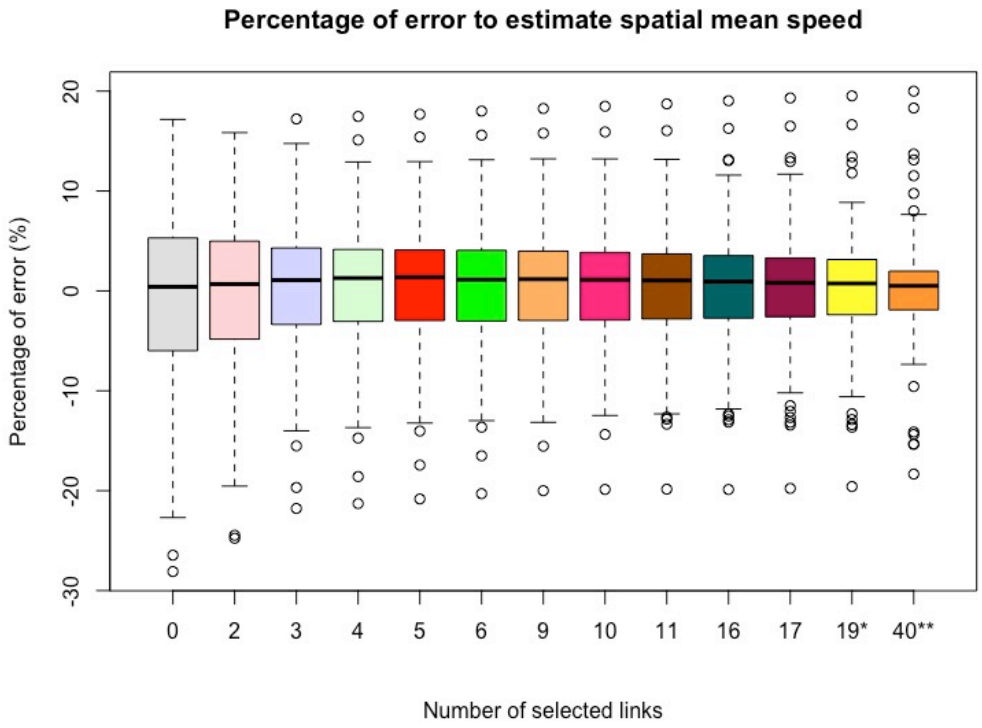



Fig. B.8 – *The percentage error of models with fewer predictors than optimized lambdas.*

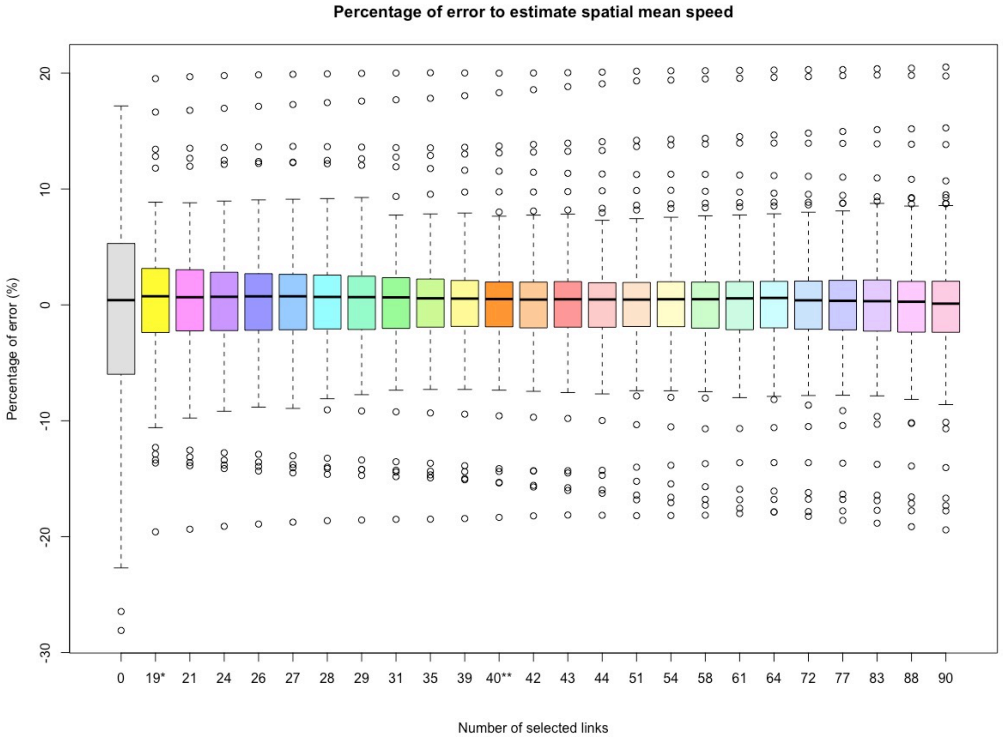
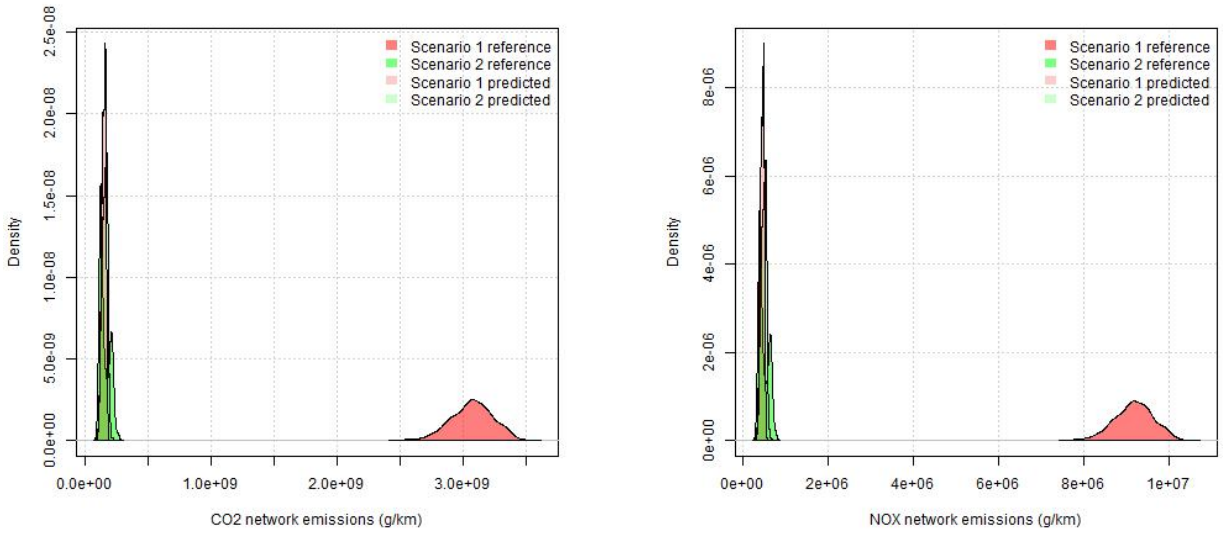



Fig. B.9 – *The percentage error of models with more predictors than the optimized lambdas.*

## B.3 Influence of route choice on the selection

The figures below present the influence of route choice using model selection fitted by LASSO. The results are shown for traveled distance, and $CO_2$ and $NO_x$ emissions. For the mean speed results see section 3.3.1.



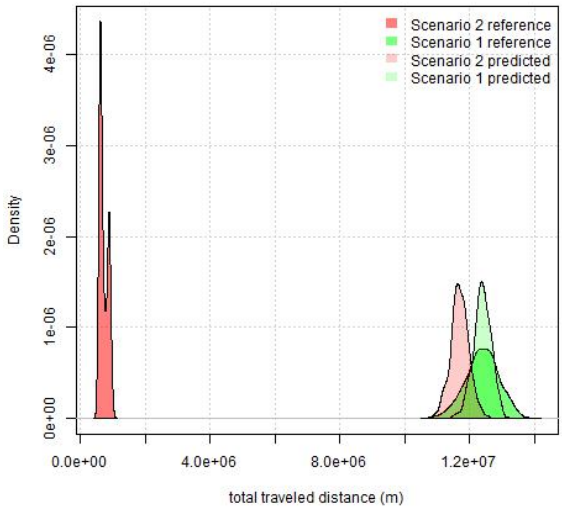

(a) $CO_2$ emission densities.

(b) $NO_x$ emission densities.

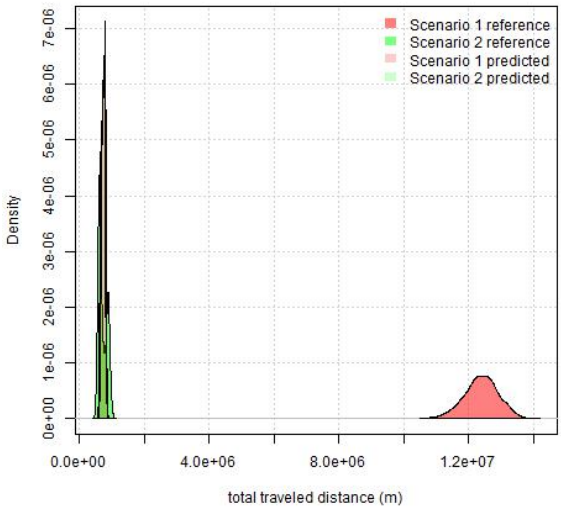
Fig. B.10 – *Pollutant emission densities using the observation of the evening peak as the training set and the morning peak as the validation set.*

Figure B.10 shows the values of $CO_2$ in $(a)$ and $NO_x$ emissions in $(b)$ using the evening data to fit a model selection using LASSO and validate the model using morning data. As explained in 3.3.1, LASSO fits a model selection around the median values of the evening data. The morning data presents more emissions than the evening, and LASSO cannot cope with this disparity using the selection method.

The same occurs for traveled distance. There are more trips in the morning than in the evening. Figure B.11 presents the traveled distances in the network of Paris. In $(a)$ the morning data was used as the training set and in $(b)$ the evening data was used to fit the selection model.




(a) Traveled distance densities of the morning and evening data and the prediction values using the morning data as the training set to fit a model.




(b) Traveled distance densities of morning and evening data and the prediction values using the evening data as the training set to fit a model.

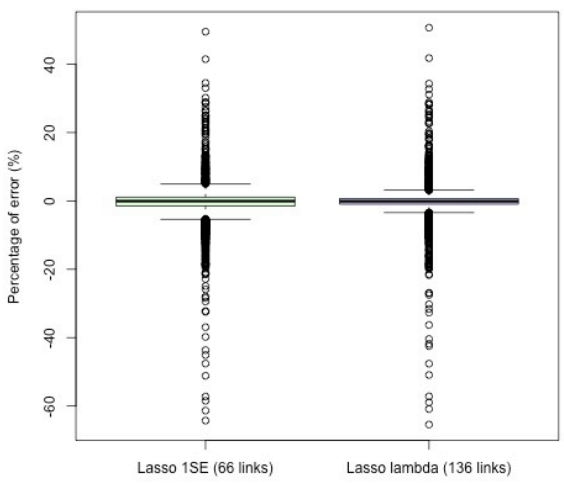
Fig. B.11 – *Traveled distance densities for both cases of the study.*

# C

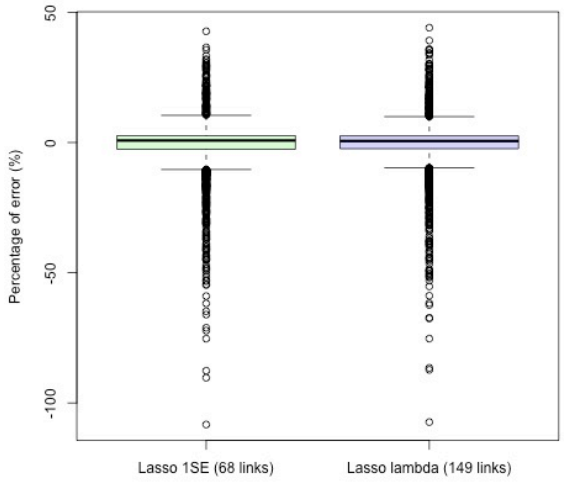
# Static/Dynamic dataset appendix

## C.1 Models proposed by LASSO

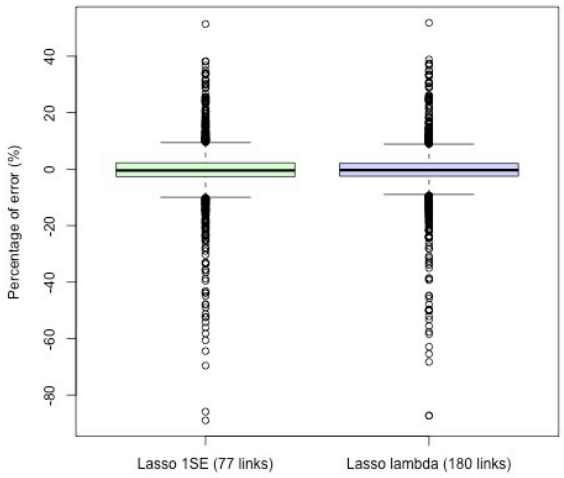
Figure C.1 shows the error distribution of both models proposed by the LASSO method. These figures are the same as those shown in 3.19, but considering their respective outliers in the error distribution.



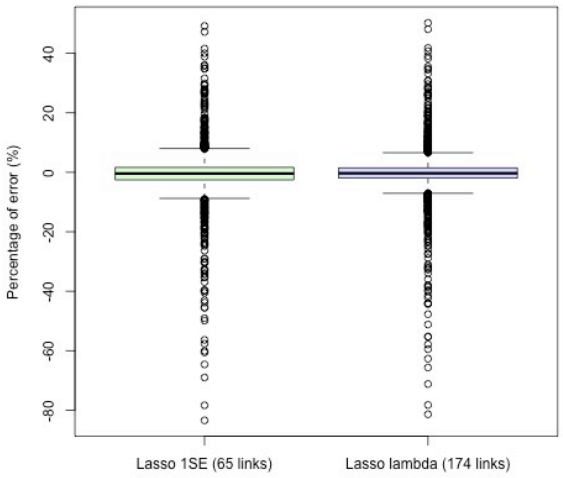

(a) Traveled distance.




(b) Spatial mean speed.




(c) $CO_2$ emissions.




(d) $NO_x$ emissions.

Fig. C.1 – *Error distribution of variables in the static/dynamic dataset.*

## C.2 Model size and error distribution

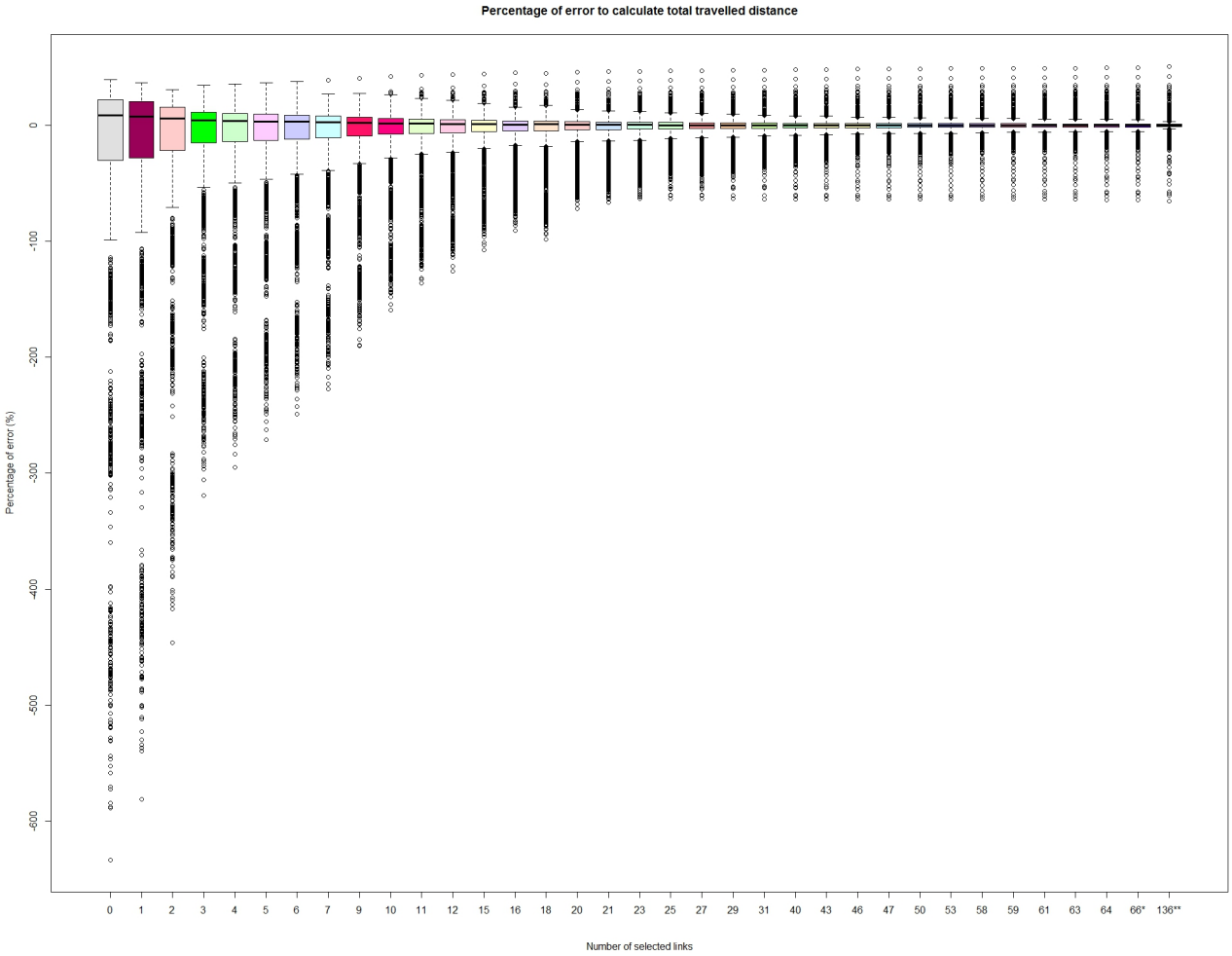



Fig. C.2 – *The error distributions of models fitted using LASSO and that have fewer predictors than the 1 SE lambda model.*

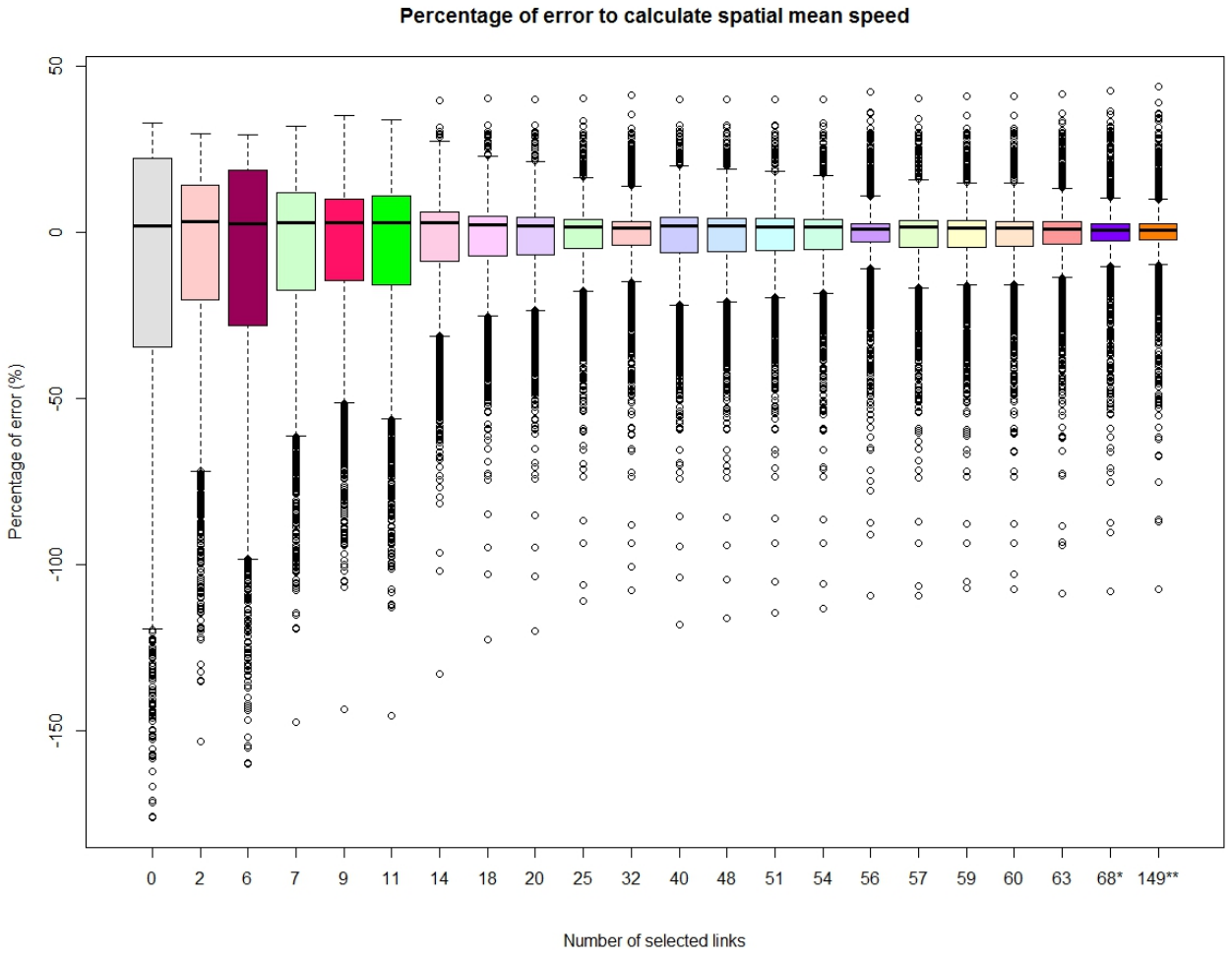



Fig. C.3 – *The error distributions for all lambdas that have models with fewer predictors than 1SE-lambda.*

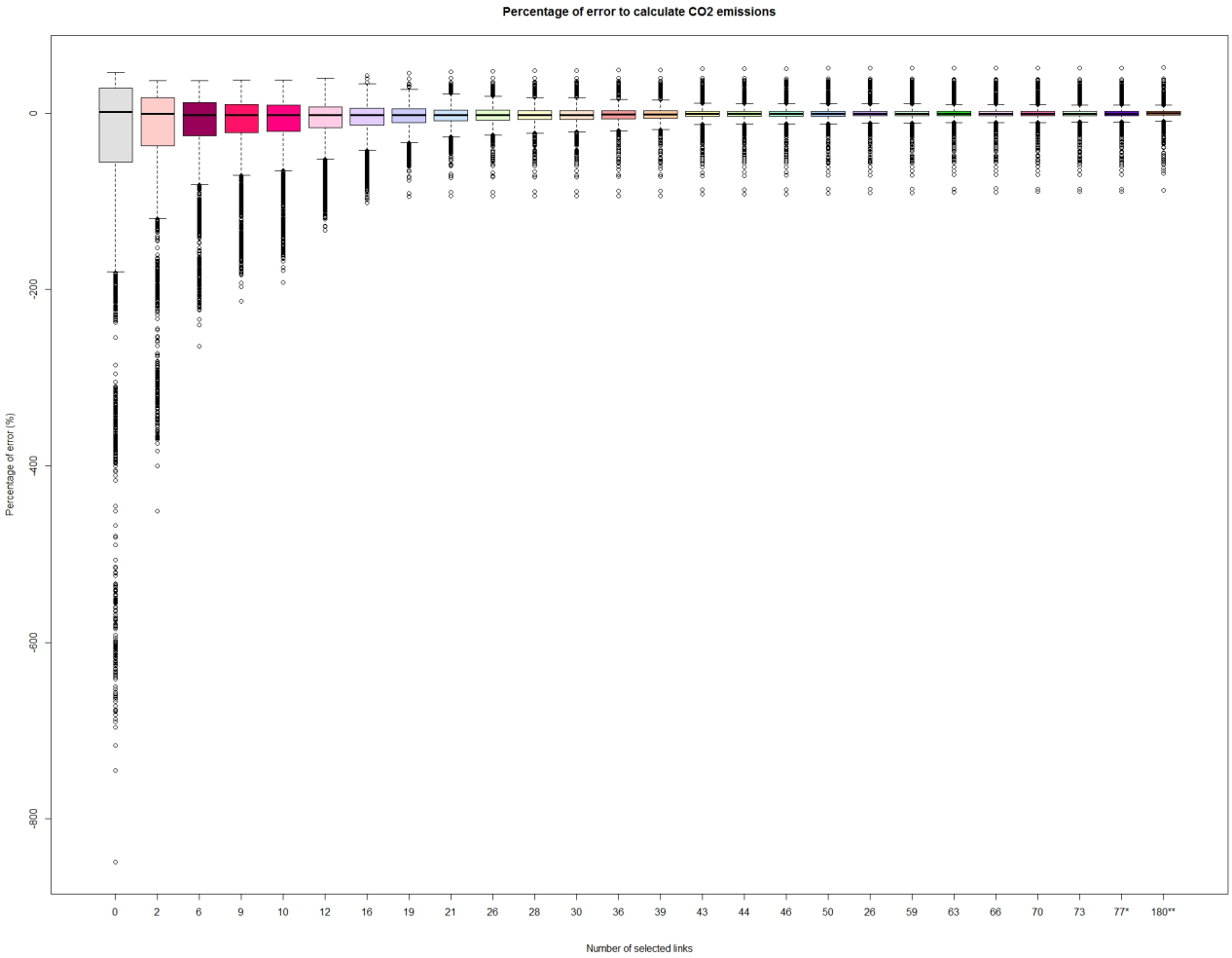



Fig. C.4 – *The error distributions from network $CO_2$ emission models fitted using LASSO with fewer predictors than 1 SElambda.*

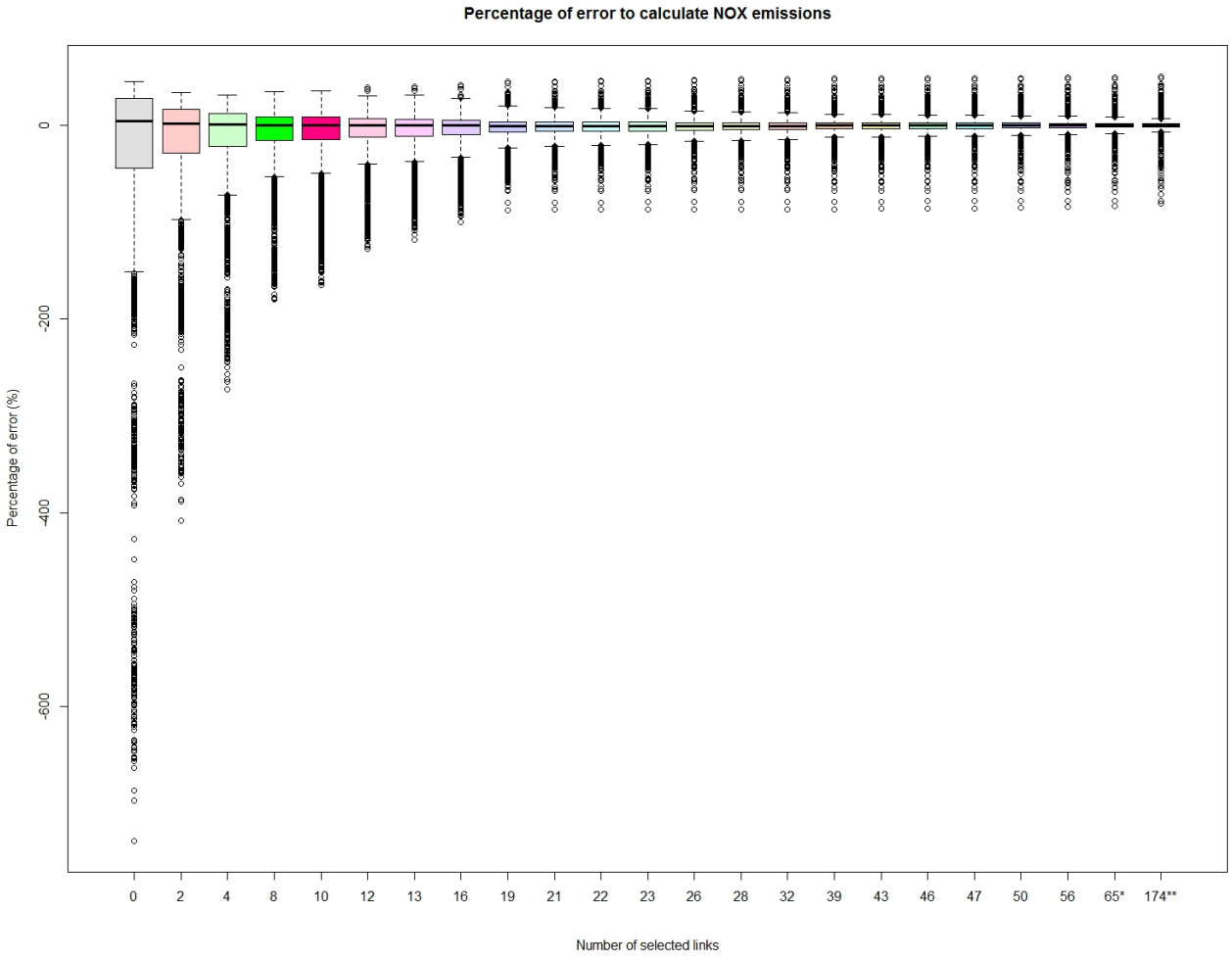



Fig. C.5 – *The error distributions from network $NO_x$ emission models fitted using LASSO and with fewer predictors than 1SElambda.*

## C.3 The robustness of the models

Table C.1 shows the mean square error of the training and validation sets. This function measures the average of the squares of the errors, i.e. the difference between the estimator and what is estimated. The mean square error is a measure of the quality of the estimator and it is always non-negative. The closer the value are to zero the better.

| VARIABLES → | Model size | | DTP (km) | | VIT (km/h) | | $CO_2$ (kg) | | $NO_x$ (kg) | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 66 | 28,70% | **0,060%** | **0,086%** | 0,090% | 0,129% | 0,088% | 0,122% | 0,076% | 0,107% |
| VIT | 68 | 29,57% | 0,059% | 0,085% | **0,084%** | **0,129%** | 0,087% | 0,119% | 0,075% | 0,107% |
| $CO_2$ | 77 | 33,48% | 0,058% | 0,084% | 0,064% | 0,129% | **0,088%** | **0,120%** | 0,074% | 0,105% |
| $NO_x$ | 65 | 28,26% | 0,058% | 0,084% | 0,084% | 0,129% | 0,088% | 0,121% | **0,077%** | **0,108%** |
| Network mean values | | | *1,07E+03* | *1,08E+03* | *15,52* | *15,56* | *2,61E+05* | *2,63E+05* | *7,85E+02* | *7,97E+02* |

Tab. C.1 – *The mean square error of all the models.*

## C.4 The best model for each variable

Figure C.6 shows the error distribution of all the models which estimate the spatial mean speed. The figure does not consider the outliers that are shown in 3.32.

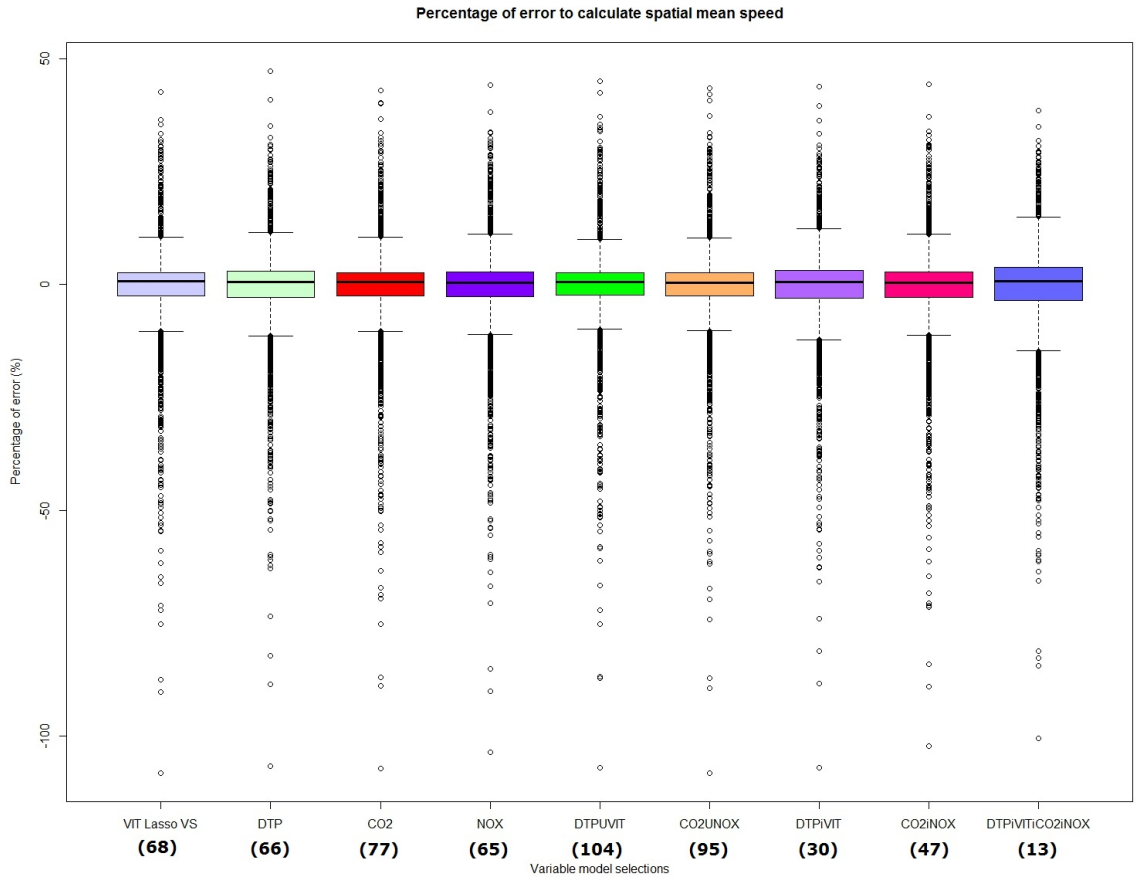



Fig. C.6 – *Error distribution from models that estimate the spatial mean speed.*

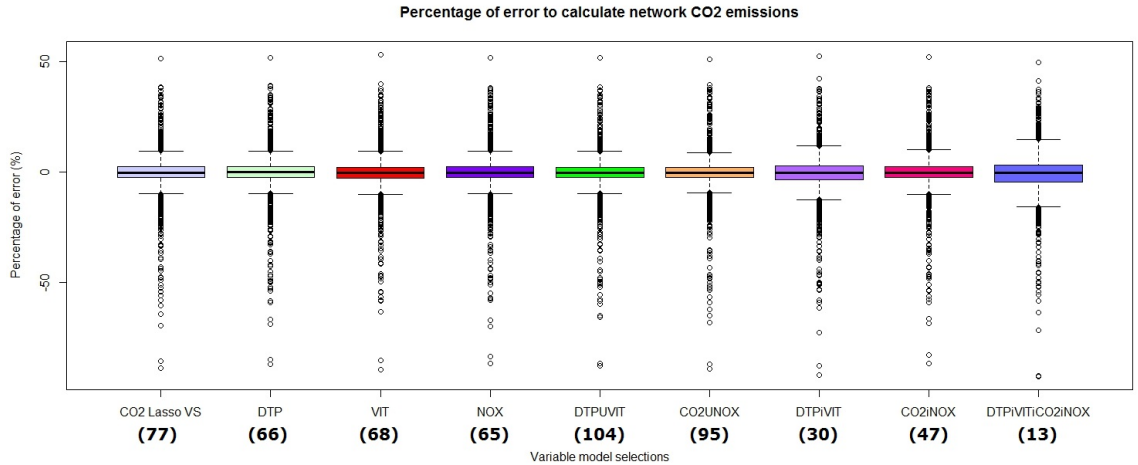



Fig. C.7 – *Error distribution of models used to determine the network $CO_2$ emissions.*

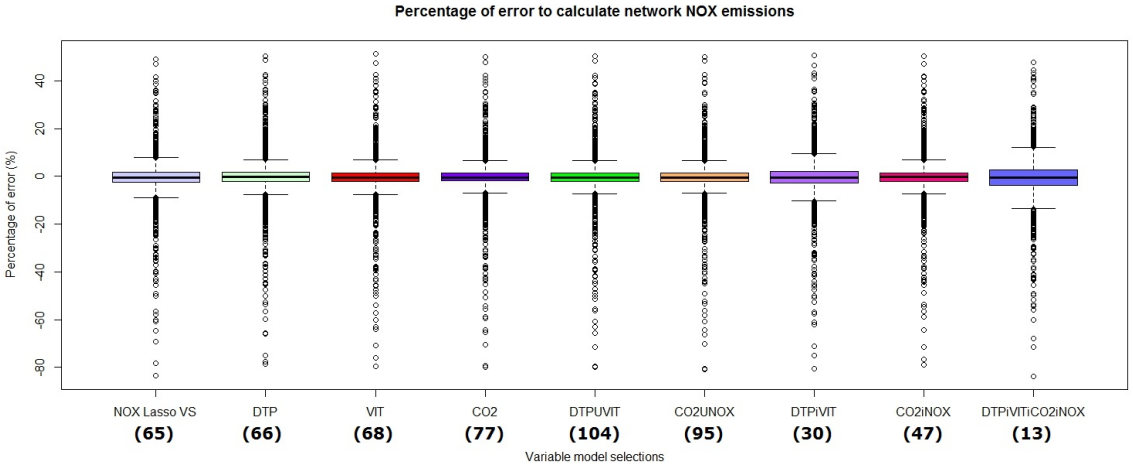
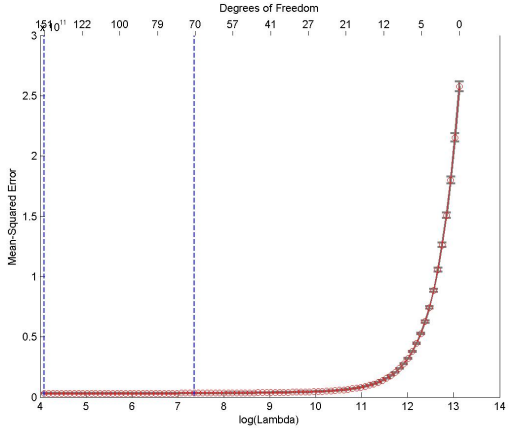


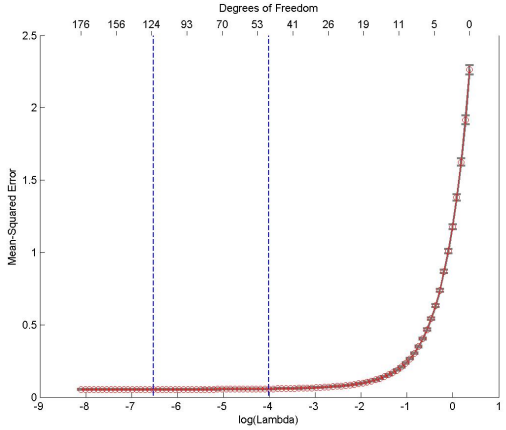

Fig. C.8 – *Error distribution of models used to estimate the network $NO_x$ emissions.*

## C.5 The influence of route choice on the selection

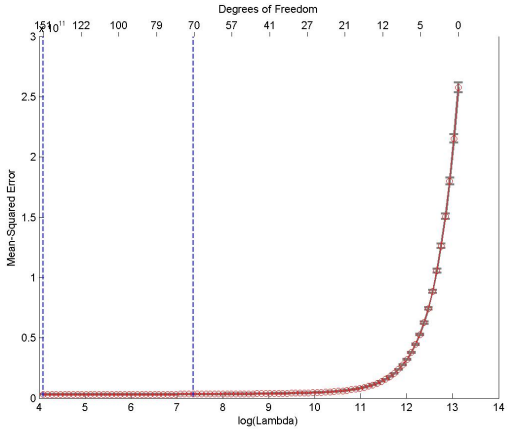
Figure C.9 shows the cross-validation for each lambda calculated by LASSO. Each variable was considered according to scenario 1 (morning peak) as the training set and scenario 2 (evening peak) as the validation set.



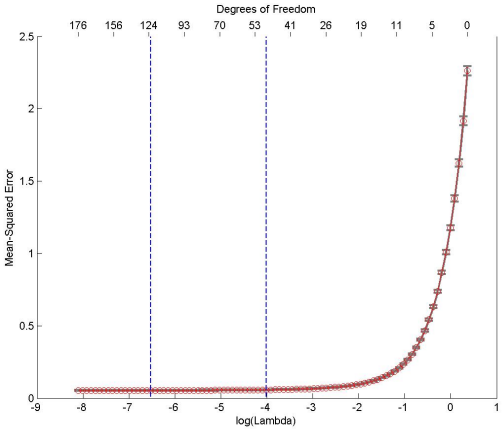

(a) Traveled distance cross-validation.




(b) Spatial mean speed cross-validation.



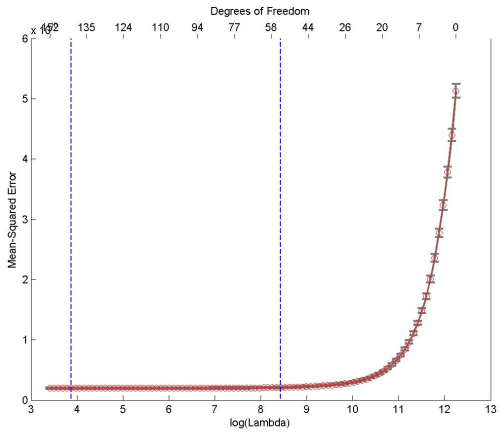

(c) $CO_2$ emission cross-validation.



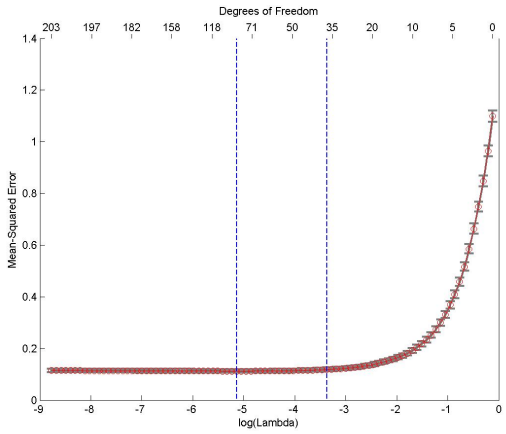

(d) $NO_x$ emission cross-validation.

Fig. C.9 – *LASSO Cross-validation of variables using morning data.*

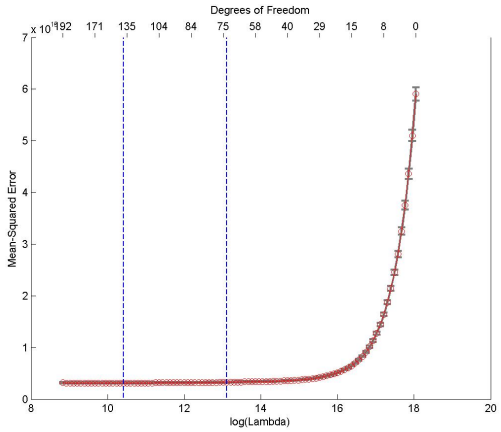
The inverse situation is proposed. Scenario 2, the evening data, will be used as the training set to build a model using the LASSO method and the model will be validated with the morning data to estimate it. Figure C.10 shows the cross-validation curves representing the two optimal models.



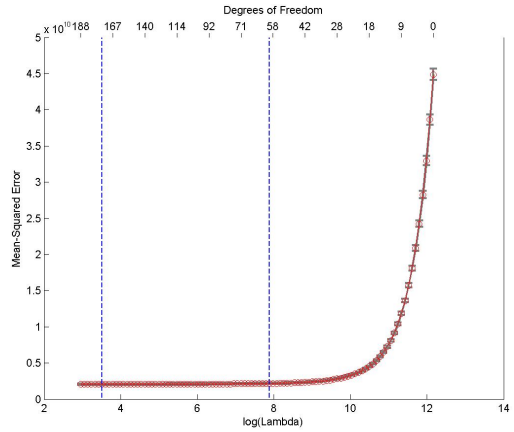

(a) Traveled distance cross-validation.




(b) Spatial mean speed cross-validation.



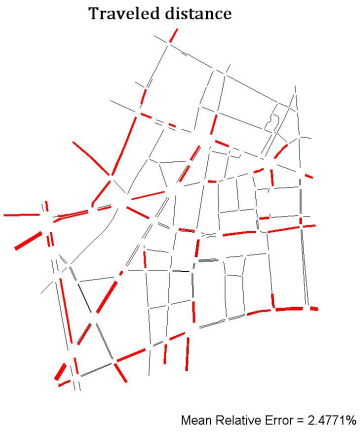

(c) $CO_2$ emission cross-validation.



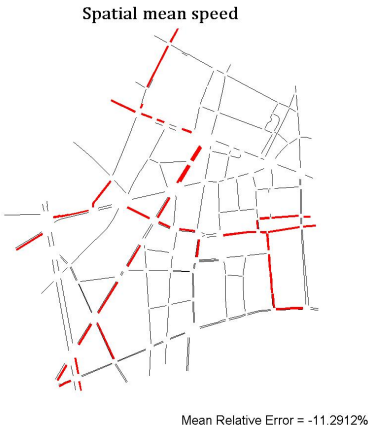

(d) $NO_x$ emission cross-validation.

Fig. C.10 – *LASSO Cross-validation of variables using evening data.*

The links selected using the evening data to fit a model and validate it using the morning data are presented for each variable in C.11. The percentage of common links between them are shown in C.2.



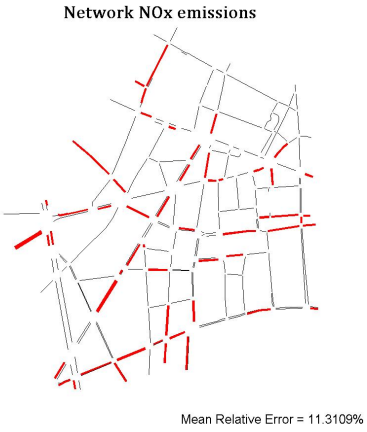

(a) Selected traveled distance links.




(b) Spatial mean speed selected links.




(c) Selected $CO_2$ emission links.




(d) Selected $NO_x$ emission links.

Fig. C.11 – *Selected links by variable using the evening data.*

| → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| DTP | 100% | 53% | 52% | 67% |
| VIT | 36% | 100% | 29% | 32% |
| $CO_2$ | 64% | 53% | 100% | 79% |
| $NO_x$ | 64% | 45% | 62% | 100% |

Tab. C.2 – *Common links between variables.*

# D

# Dynamic/Dynamic dataset appendix

## D.1 Model size and error distribution

Figure D.1 shows the error distribution for each model with a smaller size than the 1SE lambda model proposed by LASSO for the traveled distance.

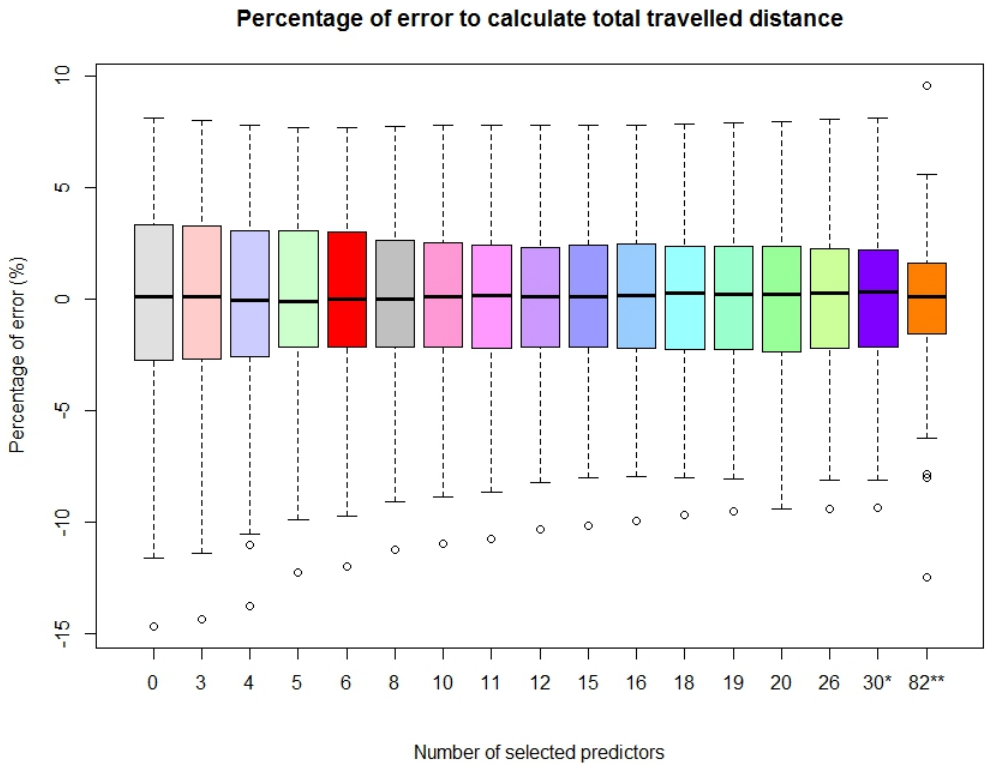



Fig. D.1 – *Error distributions versus number of predictors to estimate network traveled distance.*

Figure D.2 shows the error distributions for different model sizes smaller than the 1SE lambda model for spatial mean speed.

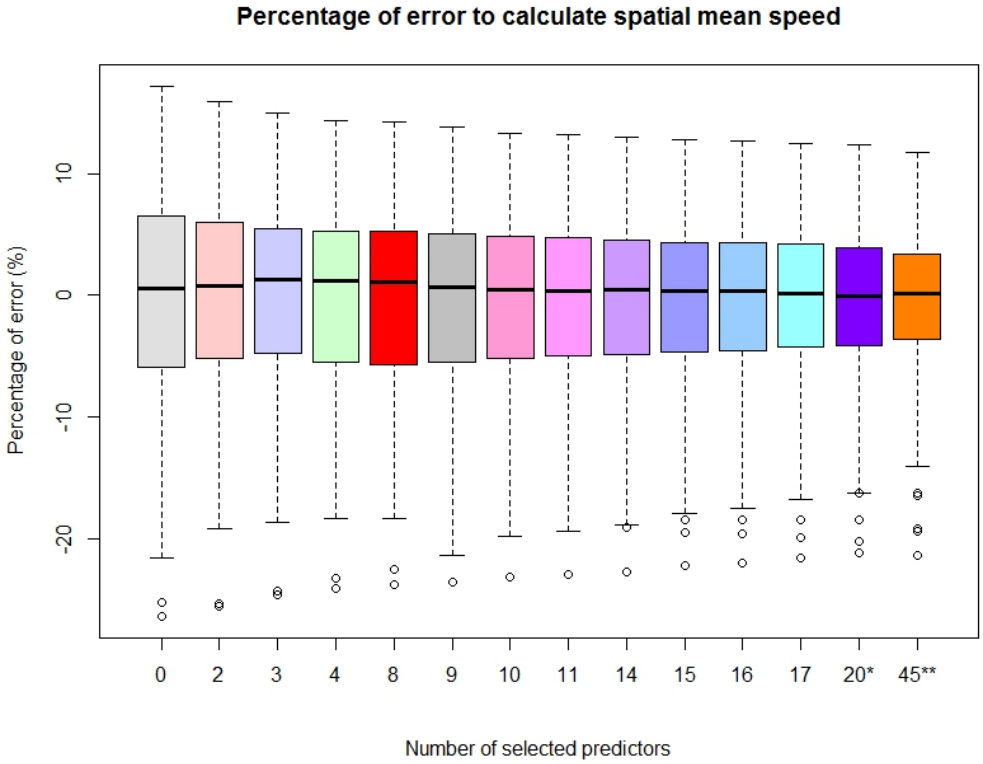



Fig. D.2 – *Error distributions versus number of predictors to estimate spatial mean speed.*

Figure D.3 shows the error distributions for different model sizes smaller than the 1SE lambda model for $CO_2$ emissions.

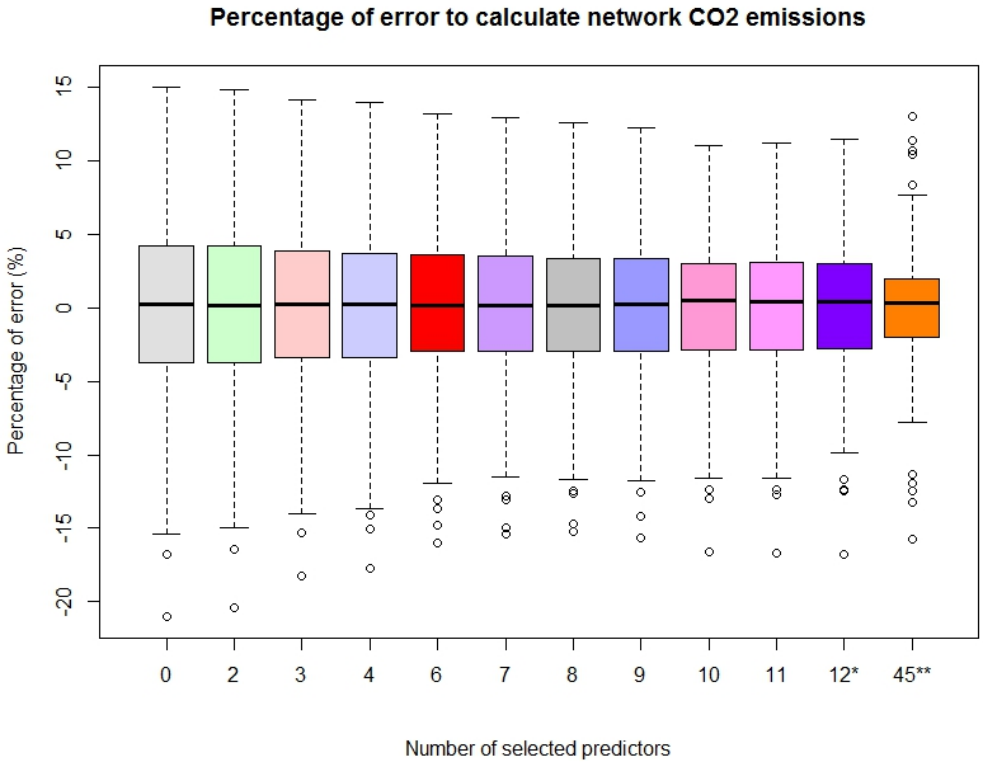



Fig. D.3 – *Error distributions versus number of predictors to estimate network $CO_2$ emissions.*

# E

## Comparison between datasets

The evolution of error by time period when the static/static (SS) and static/dynamic (SD) datasets are compared on the same temporal scale is presented here. The models built in SS are applied to the SD data. Figures E.1 and E.2 shows the evolution of errors for both pollutant emissions.

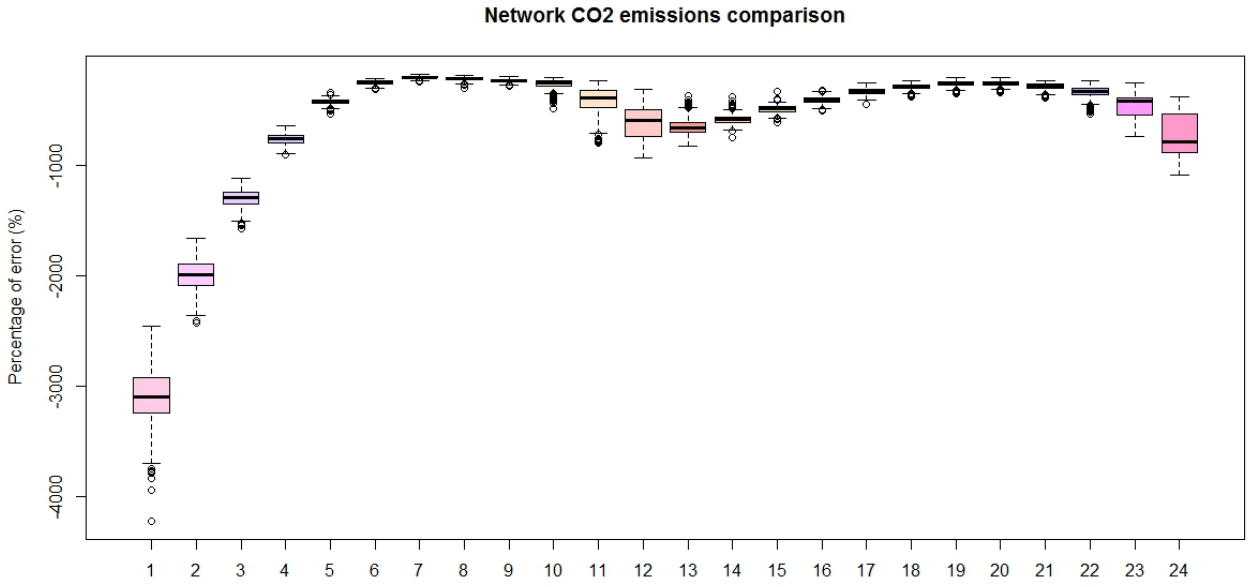



Fig. E.1 – *Error distribution by time period for network $CO_2$ emissions using the SS model applied to SD data.*

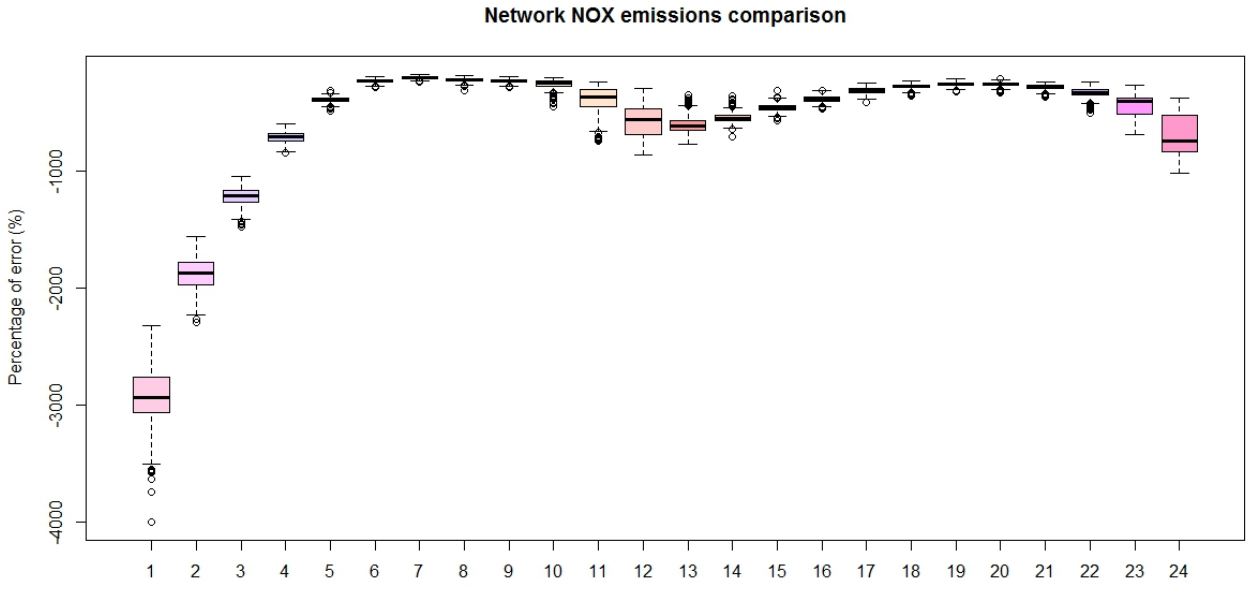



Fig. E.2 – *Error distribution by time period for network $NO_x$ emissions using the SS model applied to SD data.*

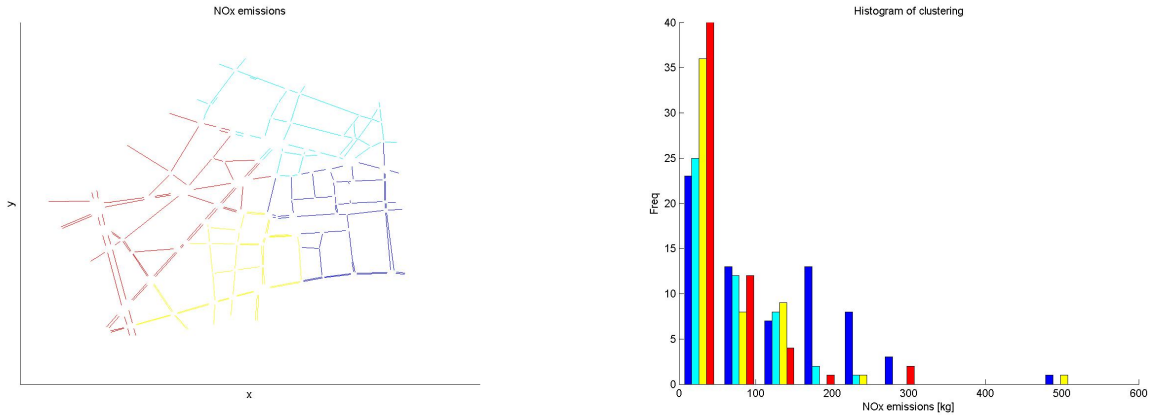
# F

# Network partitioning for the static/static dataset

## F.1 $NO_x$ network partitioning

The partitioning of the network based on the $NO_x$ emissions is explained here. Its partition into 4 and 6 clusters was similar to that of $CO_2$ emissions in the same dataset.

Partitioning into 4 clusters is shown in figure F.1.




(a) Network partitioning into 4 clusters.

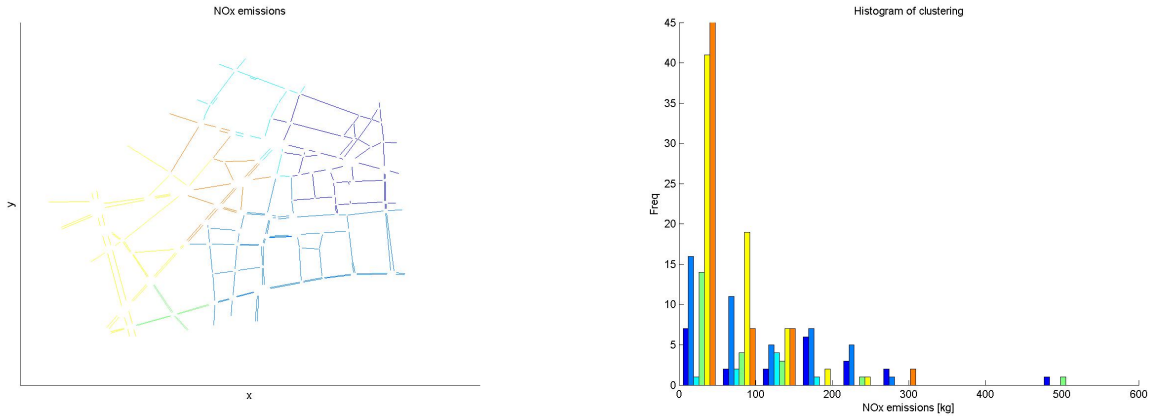
(b) Clustering histogram.

Fig. F.1 – *Network partitioning into 4 clusters based on $NO_x$ missions.*

The number of links and their contribution to the total value are the same as for the $CO_2$ partition. The red cluster contains 68 links and contributes 47.6% of the total $NO_x$ emissions to the network. The yellow cluster contains 48 links and contributes 18.4% of the emissions. The cyan cluster contains 55 links and contributes 17.2% of the total emissions. Finally, the blue cluster contains 59 links and contributes 16.8% of the total emission values at the network level. All the clusters except the yellow one contained 3 random selected links. The yellow cluster had only two links.

Figure F.2 shows the network partitioning into 6 clusters based on the daily network $NO_x$ emission values.

Like the partition into 4 clusters, the partition into 6 clusters has the same partition as that for $CO_2$. The orange cluster contains 23 links and contributes 20% of the $NO_x$ emissions. The yellow cluster contains 45 links and contributes 27.6%. The green cluster contains only 8 links that correspond to 5.5% of the contribution. The cyan cluster contains 23 links and contributes 9.8% of the emissions. The light blue cluster is the biggest with 70 links and contributes 22.2% of the emissions. The blue




(a) Network partitioning into 6 clusters.

(b) Clustering histogram.

Fig. F.2 – *Network partitioning into 6 clusters based on the $NO_x$ missions.*

cluster is based on 61 links and contributes 14.9% of the emissions. The orange, green and cyan clusters each contained only one selected link. The yellow cluster contained 2 selected links and the light blue and blue ones each contained only 3 selected links.

# G

# Lottery method applied to the static/dynamic dataset

## G.1 $CO_2$ outliers

The estimation error distribution of the $CO_2$ emissions taking into account the outliers present in the models are shown in figure G.1.

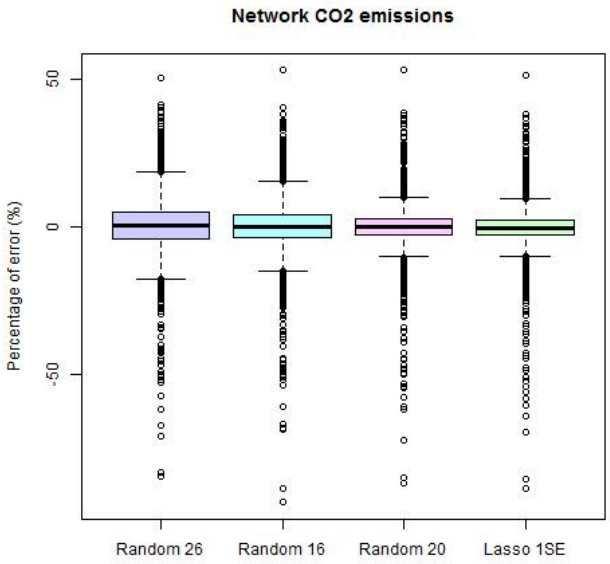



Fig. G.1 – *The $CO_2$ error distributions.*

The conclusions were given in the corresponding section. The error distributions of all the models presented are similar and their outliers reach a percentage error exceeding ±50%. The percentages of outliers are similar, around 7.3% of the data. The common links between the three cases of random draws and LASSO are between 33% and 38%, which prove that the models are different and not based on the same links.

## G.2 $NO_x$ outliers

The random draws of the $NO_x$ emissions considering the outlier distributions are shown in figure G.2.

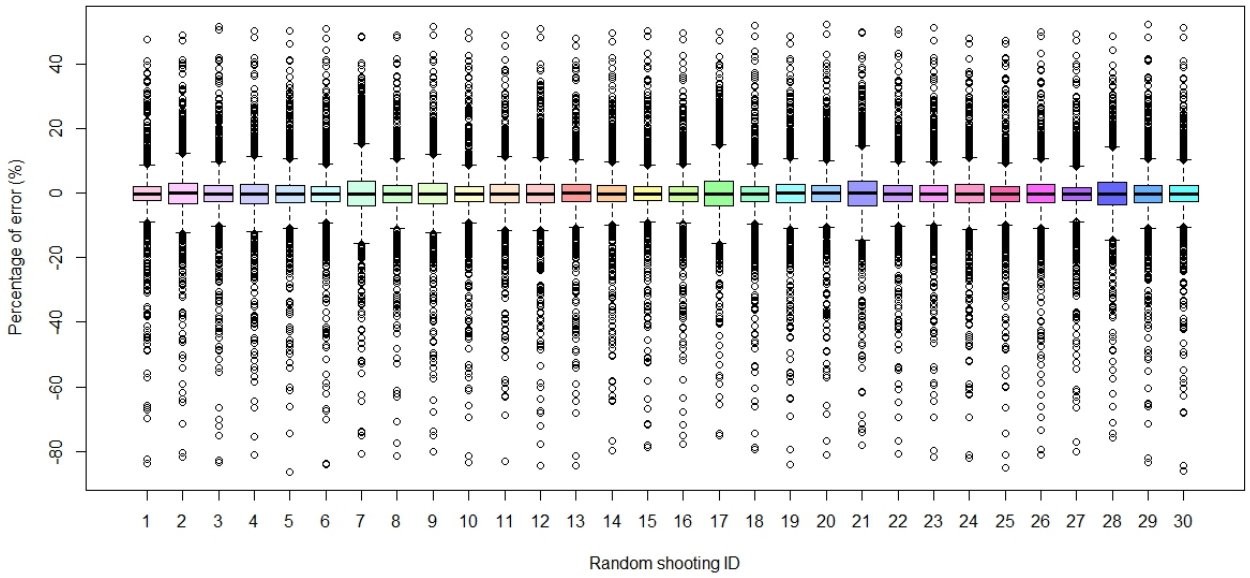



Fig. G.2 – *The $NO_x$ error estimations for all the random draws taking into account the outliers.*

The error distribution of the $NO_x$ emission estimation taking into account the outliers present in the optimal models are shown in figure G.3.

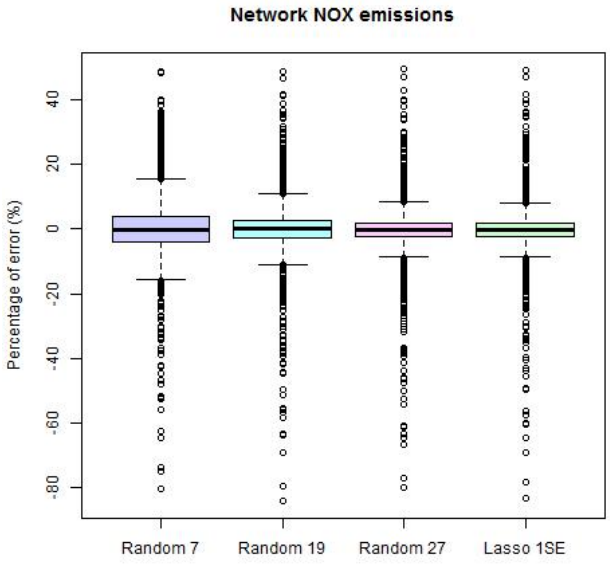



Fig. G.3 – *The $NO_x$ error distributions.*

The $NO_x$ emissions follow the same trend as that of the $CO_2$ emissions; the outlier distribution is similar between all the models and even between all the random draws. The outliers represent between 7% and 8% of the data and they can reach a percentage error exceeding -80% and +40%. The error distributions of 92% of the data are shown in figure 4.15.

# H

## Network partitioning for the static/dynamic dataset

### H.1 $CO_2$ emissions

Figure H.1 shows the estimation errors of the model fitted to the random selection in the clusters compared to the optimal model of LASSO considering the outlier distributions.

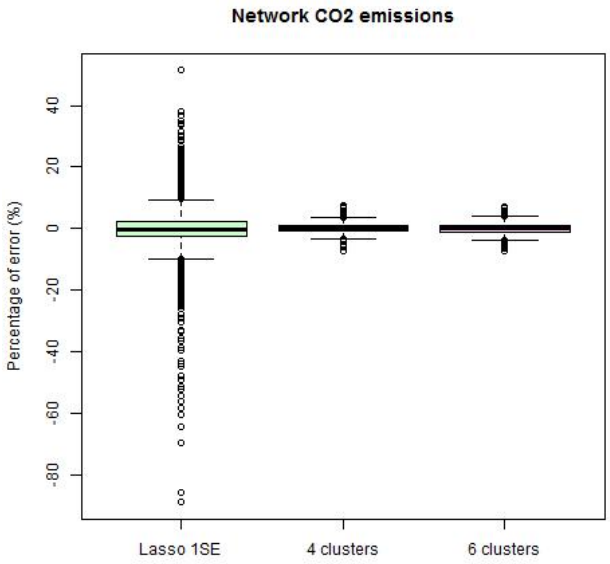



Fig. H.1 – $CO_2$ *error distributions.*

The conclusions were made in the corresponding section. In general, clustering the network vastly reduces the number of outliers and their estimation errors, which means that more traffic situations can be estimated by the model.

## H.2 $NO_x$ emissions

Figure H.2 shows the estimation errors from the model fitted to the random selection inside the clusters compared to the optimal LASSO model, considering the outlier distributions.

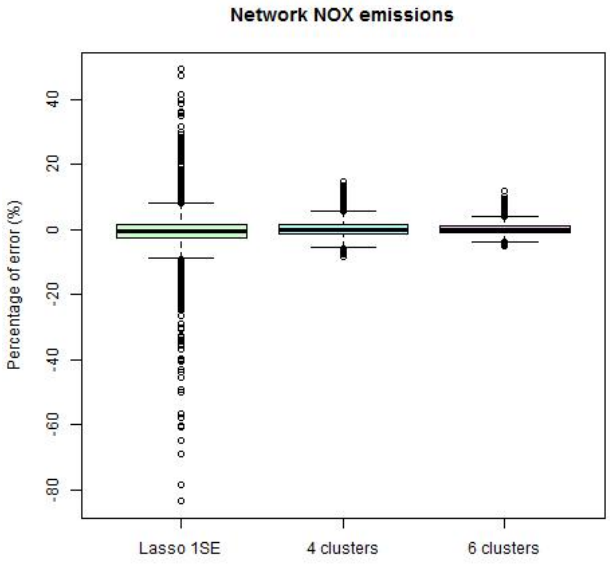


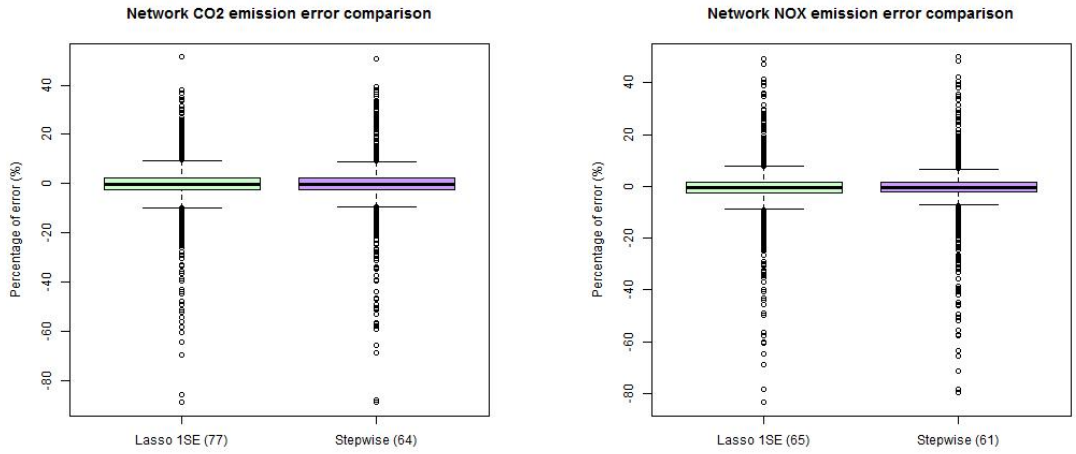

Fig. H.2 – *$NO_x$ error distributions.*

The $NO_x$ emissions present slightly more outliers in both predictive models, the 1SE lambda model has 7.9% outliers compared to the ranked links with 8.2% outliers.

# I

# Stepwise selection

In this section the results from the stepwise models in the static/dynamic dataset are presented. The error distributions for both pollutant emissions taking into account the outliers are presented here and the conclusions are given in 4.3.2. Figure I.1 shows the error distributions for both pollutants.




(a) $CO_2$ error distribution taking into account outliers.

(b) $NO_x$ error distribution taking into accout outliers.

Fig. I.1 – *Comparison of associated errors from the LASSO and the stepwise model.*

# J

# Comparison of all methods

## J.1 Static/dynamic datasets

The estimation errors of all the models fitted to the estimated $CO_2$ emissions, taking into account the outliers, is shown in figure J.1.

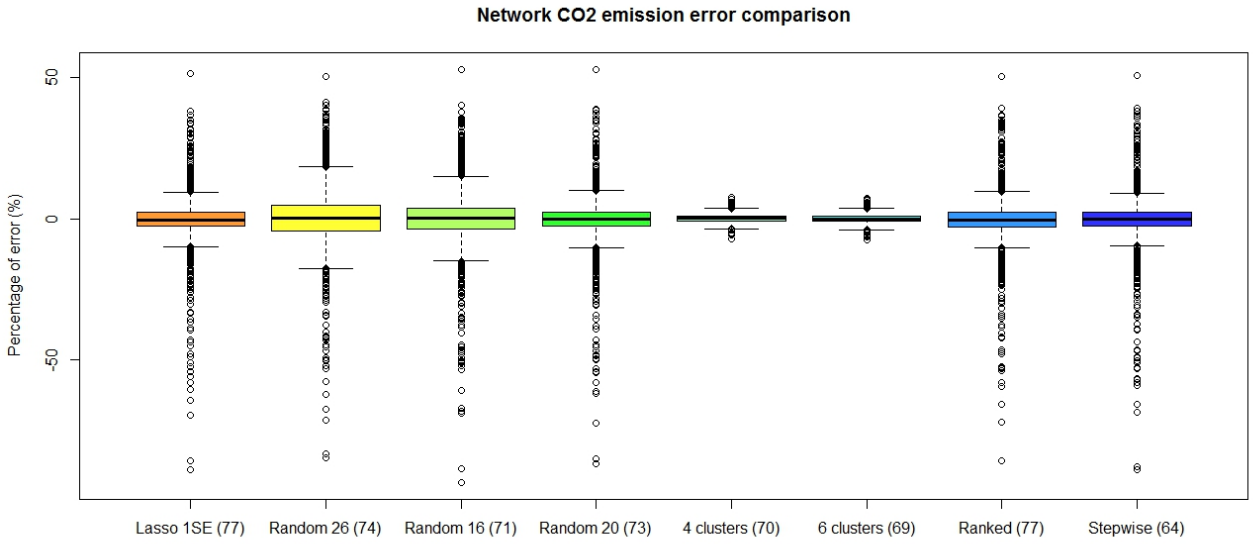



Fig. J.1 – $CO_2$ *error distributions.*

The estimation error from all the models fitted to estimate the $NO_x$ emissions taking into account the outliers is shown in figure J.2.

The conclusions are given in the respective section. Of all the methods, the clustering method vastly reduced the quantity of outliers and the error distributions of the resulting models.

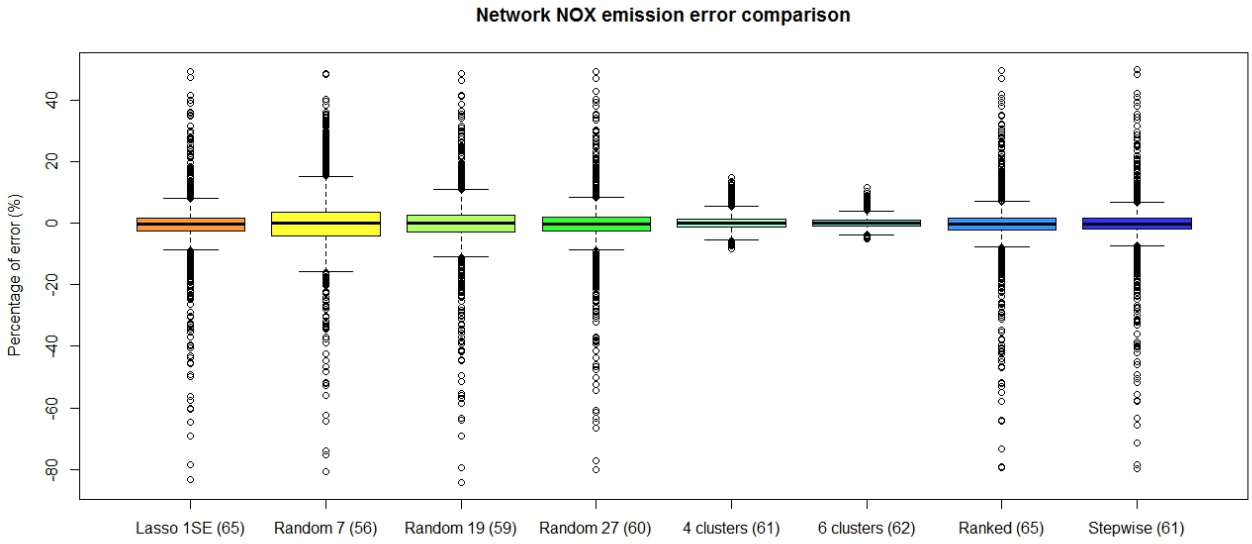



Fig. J.2 – *$NO_x$ error distributions.*




N° d'ordre NNT : 2017-LYSET-008

**THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON**

Opérée au sein de

**l'Ecole Nationale des Travaux Publics de l'Etat**

**École Doctorale** N° 162

**MEGA (Mécanique, Energétique, Génie Civil et Acoustique)**

**Spécialité / discipline de doctorat : Génie Civil**

soumise en vue d'une soutenance publique le 09 Octobre 2017, par :

**Nicole SCHIPER**

---

# Échantillonnage des données de trafic pour l'estimation de la pollution atmosphérique aux différentes échelles urbaines: Résumé en français

---

Devant le jury composé de :

| | | |
|---|---|---|
| LATIFA OUKHELLOU, | Directrice de recherche (Université de Marne-la-Vallée) | *Rapporteur* |
| NIKOLAOS GEROLIMINIS, | Associate Professor (École polytechnique de Lausanne) | *Rapporteur* |
| CHRISTINE SOLNON, | Professeur (Université de Lyon) | *Présidente* |
| DELPHINE LEJRI, | Ingénieur des TPE, Docteur (Université de Lyon) | *Encadrante* |
| LUDOVIC LECLERCQ, | Directeur de Recherche (Université de Lyon) | *Directeur de Thèse* |

Thèse préparée au LICIT (Laboratoire d'Ingéniérie Circulation Transport)

# Introduction

Ce document résume très brièvement le manuscrit de thèse intitulé "Échantillonage des données de trafic pour l'estimation de la pollution atmosphérique aux différents échelles urbaines" présenté à l'Université de Lyon pour l'obtention du titre de docteur en Génie Civil.

Le travail de recherche présenté dans ce document a été réalisé dans le cadre d'une collaboration entre l'ENTPE/Université de Lyon et IFSTTAR. Cette thèse s'inscrit dans le projet de recherche de l'équipe AMMET (Analyse et Modélisation Multi-Échelles du Trafic) du LICIT dans son composant couplage multiphysique pour l'évaluation environnementale des externalités du trafic. Le sujet de recherche que je présente se concentre sur l'échantillonnage des données de trafic. Il s'agit de déterminer des méthodes efficaces pour identifier l'échantillon de taille minimal permettant de déterminer avec un niveau de précision suffisant les caractéristiques des émissions de polluants dans la population totale (moyenne, total, écart type,...). Dans le cadre de cette thèse on travaillera dans un premier temps sur une population Eulériennes (capteurs) parfaitement connue car issue d'une simulation dynamique microscopique à grande échelle.

Par rapport à l'objectif de cette thèse, les principaux verrous sont liés à la prise en compte des corrélations spatio-temporelles entre les informations trafic et plus généralement à la prise en compte de la dimension temporelle. Les corrélations spatio-temporelles sont liées aux ondes de congestion et de modifications de la demande qui se propagent dans le réseau. Ainsi, il est important de définir une méthode permettant de prendre en compte ces corrélations pour la segmentation de la population pour définir un échantillon représentatif. De plus, la dynamique du trafic modifie les conditions de circulation sur différents horizons temporels (de la seconde à la journée). Il est important pour estimer correctement les émissions de polluants mais également leurs chroniques temporelles de prendre en compte ce facteur temps. Cela pourra notamment conduire à un échantillonnage différentié suivant les périodes de la journée. Pour toutes ces études, une bonne connaissance du couplage entre les modèles de trafic et d'émissions de polluants nous sera profitable.

Dans ce résumé, les sections sont introduites séparément. les principales points et le résultats de l'études sont présentes. Cela permet d'avoir une vision globale de ce qui a été réalisé et des conclusions qui ont été faites. Le travail de thèse est structuré autour de cinq chapitres. Le chapitre 1 présente une revue de la littérature existante sur différents modèles de trafic et d'émission et les différentes approches de couplage entre ces modèles. Le chapitre 2 présente une analyse de la sensibilité sur la précision de la représentation du trafic en fonction de la source d'information et de son incidence sur les estimations des émissions dans les zones urbaines et comprend également l'analyse descriptive des données de trafic et d'émissions du réseau utilisé dans le cadre de cette thèse et leurs corrélations. Les chapitres 3 et 4 étudient différentes méthodes d'échantillonnage pour estimer le trafic et les émissions dans différents niveaux spatio-temporels. Enfin, le chapitre 5 présente l'analyse des incertitudes inhérentes au modèle d'émission et de la répartition dans les estimations d'émission dans différentes échelles spatio-temporelles.

# 1

# Estimation des émissions routières à partir des modèles de trafic

## 1.1 Transport et environnement : les émissions des véhicules

Les émissions de polluants sont une préoccupation majeure à travers le monde et ont augmenté à un taux relativement important au cours des 25 dernières années. De nombreux secteurs industriels participent à cette croissance, y compris l'énergie, les procédés industriels, l'agriculture, les déchets, les solvants et d'autres produits. (DOE, 2015) montre que dans les émissions de transport sera le seul secteur parmi d'autres qui présente une tendance qui augmentera au lieu de se stabiliser ou diminuer dans un proche avenir compte tenu d'une tendance de croissance actuelle et prévue dans les transports.

Selon (EEA, 2017a), les voitures particulières dans la zone EU-28 ont eu une augmentation moyenne de 1,2% par an entre 2000 et 2013, ce qui représente de 415 à 490 voitures par 1000 habitants. Cette croissance est la conséquence d'autres facteurs importants tels que : (i) un nombre décroissant de personnes par ménage, (ii) un nombre croissant de voitures par famille et (iii) une augmentation de la distance moyenne de déplacement parce que l'accès aux transports publics est moins élevé. Dans cette composition de la flotte des véhicules, le nombre de voitures diesel a augmenté de 27% à 38% pour la même période en Union Européenne (UE) et le pourcentage de voitures diesel en France est particulièrement élevé et peut atteindre 68% la flotte. En ce qui concerne les émissions de polluants, le transport de véhicules légers est responsable d'environ 85% de l'émission totale des routes (IPCC, 2013). Certaines villes de l'UE présentent une augmentation considérable des oxydes d'azote ($NO_2$) et les matières des particules fines ($PM_{2,5}$ et $PM_{10}$) mesurées proches du trafic sont solides liées à un niveau d'air de mauvaise qualité et la source provient principalement des voitures diesel. L'Organisation mondiale de la santé (OMS) a fixé les valeurs des lignes directrices sur la qualité de l'air pour la protection de la population humaine dans la zone EU-28 et a conclu que plus de 90% de la population sont exposés à un polluant au moins à des niveaux dangereux pour la santé. Entre 2006 et 2014, la zone de population urbaine de l'UE-28 a été exposée à une concentration en excès en fonction des limites cibles imposées par (WHO, 2013) et (EU, 2013) : (i) pour les particules fines ($PM_{2,5}$) 12.5%, en moyenne, de la population urbaine ont été exposés aux valeurs cibles de l'UE et 91%, en moyenne, ont été exposés à des concentrations supérieures aux valeurs indicatives de l'OMS ; (ii) pour les particules fines ($PM_{10}$), l'exposition respective était de 29% pour les valeurs limites de l'UE en moyenne et de 50 à 92% pour les directives de l'OMS ; (iii) pour le dioxyde d'azote, les estimations étaient de 7 à 31% en deux valeurs limites.

Compte tenu des émissions de gaz à effet de serre (GES) du transport, les émissions de dioxyde de carbone ($CO_2$), de méthane ($CH_4$) et d'oxydes nitreux ($N_2O$) ont augmenté de 2% en 2015 par rapport à 2014 (EEA, 2017b) et plus de 19% par rapport aux niveaux de 1990. La même année, le transport routier était responsable de près de 73% des émissions totales de gaz à effet de serre des transports, et de ces émissions jusqu'à 44% provenaient de voitures de passagers et 18% étaient des véhicules au poids lourds. Compte tenu des changements climatiques, le dioxyde de carbone ($CO_2$) publié par les sources de trafic est considéré comme l'un des principaux gaz à effet de serre ayant un fort impact sur le changement climatique (IPCC, 2013) et représente une menace économique, sociale et environnementale.

Compte tenu de toutes ces préoccupations, de nombreuses politiques ont été mises en oeuvre pour réduire l'exposition de la population et par conséquent augmenter la qualité de l'air. Un plan de gestion local et régional est établi pour améliorer la qualité de l'air, y compris notamment des initiatives telles que les zones à faibles émissions dans les villes. À grande échelle, les changements dans les lois et les événements politiques internationaux ont été les premières stratégies pour atténuer la pollution atmosphérique. Afin d'éviter les impacts du changement climatique, certains pays ont convenu de coopérer en vue de limiter l'augmentation de la température mondiale et les changements climatiques qui ont abouti à la signature de la Convention-Cadre des Nations Unies sur les Changements Climatiques (CCNUCC). L'objectif de la CCNUCC est d'exiger des membres des inventaires précis et réguliers des GES pour empêcher les interférences dangereuses de l'homme dans le système climatique (UNFCCC, 2014). Au niveau international, le protocole de Kyoto est le principal instrument pour atténuer les émissions de GES. Il a commencé en 1997 et a réglé des objectifs pour réduire les émissions dans les pays membres. La première période d'action du protocole de Kyoto a débuté en 2008 et s'est terminée en 2012 et la deuxième a débuté en 2013 et se terminera en 2020. Entre-temps, l'UE a décidé son propre objectif d'atténuation du changement climatique pour 2020 en réduisant 40% d'ici 2030 et 80% d'ici 2050 par rapport aux niveaux de 1990 (UNFCCC, 2012).

Pour atteindre les objectifs donnés par le lois et des accords, certaines actions ont été réglées à l'échelle locale. Au cours de la dernière décennie, des normes plus strictes ont été introduites pour accroître l'efficacité des actions. Quelques exemples d'action d'atténuation des émissions dans les pays européens sont les suivants : installer des filtres catalytiques à particules diesel dans des véhicules anciens et des autobus plus anciens pour réduire les émissions de $NO_x$ dans les zones urbaines (Carslaw and Beevers, 2005) ; La *Low Emission Zone* mise en place en 2008 à Londres qui a restreint l'entrée du véhicule poids lourds le plus polluant (HGV) dans des régions stratégiques (TfL, 2014) ; Déployer un plan de nouveaux véhicules hybrides et à faibles émissions (EURO IV), une stratégie de qualité de l'air également mise en oeuvre à Londres en 2010 et est utilisée comme modèle pour d'autres villes européennes (GLA, 2010). D'autres solutions de rechange ont été mises en oeuvre en raison de l'utilisation de carburants renouvelables dans les transports et de l'incitation à augmenter les véhicules à carburant de remplacement en proportion de la flotte totale dans le but d'atteindre 10% des énergies renouvelables dans les transports d'ici 2020 en comparaison de 5,4 % en 2013 (EEA, 2016).

De nos jours, il est préoccupant que ces stratégies ne répondent pas comme prévu ou ne suffisent pas à diminuer les émissions. (Font and Fuler, 2016) a étudié l'impact des politiques visant à réduire les émissions liées à la circulation à Londres pendant une période de 5 ans et a conclu que, malgré la réduction de la voiture et des taxis, les niveaux de polluants n'ont pas diminué en conséquence, principalement observés pour le gaz carbonique. Pour les grandes villes ou en développement, la gestion de la mobilité est un facteur important qui modifie la façon dont le véhicule est utilisé pour augmenter la capacité et l'efficacité du système de transport et, par conséquent, réduire les émissions des véhicules. De plus, cela contribue à améliorer les flux de trafic, réduisant ainsi la congestion et réduisant ainsi les émissions. Le défi consiste donc à améliorer la qualité de l'air en tenant compte de la demande croissante sans compromettre la mobilité de la population.

L'impact des mesures de contrôle des transports sur les émissions est généralement mesuré en termes de réduction des émissions de véhicules provoquées par ces stratégies. À l'heure actuelle, bon nombre des modèles de transport intègrent des technologies pour mesurer les polluants du trafic routier, afin d'aider à l'évaluation des stratégies de transport en tenant compte de leurs impacts environnementaux respectifs.

## 1.2 Le couplage entre les modèles de trafic et d'émission

Généralement, en raison de la difficulté de mesure ou des incertitudes liées à de nombreux scénarios de transport, les émissions des véhicules sont estimées en combinant les modèles d'émissions et de trafic. Par conséquent, ce couplage est généralement utilisé pour évaluer l'efficacité environnementale des stratégies de circulation et leur potentiel avant la mise en oeuvre (Jie et al., 2013).

Les modèles de trafic et d'émission ont généralement trois niveaux de représentation spatiale : macroscopique (flux de trafic), mésoscopique (groupe de véhicules) et microscopiques (véhicules in-

dividuels). Dans la représentation temporelle, ces modèles peuvent être classés comme statiques et dynamiques. Les modèles statiques supposent essentiellement que l'état de la circulation ou les émissions sont stationnaires pendant la période analysée. Ils considèrent les mouvements de la circulation et non la façon dont ils sont réalisés. Les modèles dynamiques décrivent le flux de trafic et représentent donc les situations de trafic rencontrées par les véhicules pendant leur voyage (Cappiello, 2002).

L'estimation des émissions du transport routier nécessite une information complète sur les caractéristiques du trafic en tant que composition du parc automobile et conditions de circulation. Dans ce contexte, il est nécessaire d'utiliser des modèles appropriés pour estimer avec précision le trafic et les émissions afin de représenter de près la réalité. Les modèles d'émissions statiques sont associés à des modèles de trafic statique appliqués à des études à grande échelle. L'un des modèles de trafic les plus utilisés pour les entrées générées pour les modèles d'émission, parfois les entrées agrégées dans l'espace et le temps, est le modèle de trafic statique. Ils sont largement utilisés car ils peuvent être utilisés efficacement dans les grandes zones urbaines avec un faible effort de calcul (Tsanakas et al., 2017).

L'étude menée par (Tsanakas et al., 2017) a montré que les modèles de trafic statique ne peuvent pas reproduire les phénomènes dynamiques du transport et qui peuvent conduire à une sous-estimation de 40% des émissions polluantes. Bien que de nombreuses méthodes soient disponibles et soient bien fondées et utiles pour mesurer les émissions, leur concentration et également évaluer les stratégies pour l'atténuer, cette approche peut camoufler une grande hétérogénéité de l'impact des politiques dans les zones urbaines où le trafic Les émissions apparentées ont une grande variabilité spatiale et temporelle. Ce fait souligne la nécessité de procéder à des mesures plus détaillées du trafic, des émissions polluantes et de la qualité de l'air, alliées à des méthodologies adaptées et les signaler au processus d'élaboration des politiques afin de renforcer les paquets de politiques et d'assurer des avantages pour la qualité de l'air (Frecht et al., 2015). Une bonne compréhension de la dynamique du trafic est fondamentale pour aider à choisir la stratégie d'étude la plus efficace à adopter pour chaque type de problème traité. Dans ce contexte, les simulations microscopiques peuvent reproduire l'effet que tout changement aura sur le trafic, en prédisant leur comportement, peut contribuer à définir les stratégies appropriées à adopter pour améliorer le trafic en question (Schiper et al., 2016). Par la suite, les modèles d'émissions doivent être sensibles aux effets de la dynamique du trafic sur les estimations des émissions de trafic. Dans cette thèse, nous nous concentrons sur les modèles microscopiques de trafic car c'est à l'échelle locale, surtout dans les zones urbaines, que la dynamique du trafic influence le plus l'estimation des émissions.

L'objectif des modèles d'émissions est de quantifier les émissions des véhicules en fonction du mode de fonctionnement des véhicules concernés. De tels modèles ont également été utilisés pour déterminer la quantité d'émissions générées par le trafic en raison de la difficulté de les quantifier dans le monde réel. Ils sont développés en utilisant des données de mesure sur les taux d'émission obtenus à partir des véhicules. Il existe plusieurs méthodes disponibles pour quantifier les émissions polluantes. Selon (Sturm et al., 1996), la définition du type de modèle d'émission dépend beaucoup du besoin et de la précision spécifiques requis pour décrire le comportement des émissions du trafic routier.

Les modèles d'émission sont utilisés pour deux types de calcul. Ils peuvent être utilisés pour prédire les valeurs absolues de la pollution (inventaires), comme l'identification des rues qui dépassent les normes de qualité de l'air, mais pour ce type d'analyse, un degré élevé de précision concernant les facteurs d'émission est nécessaire. Les modèles d'émission sont également utilisés pour les évaluations d'impact, comme la comparaison de différentes stratégies de trafic pour réduire les émissions. Dans ce type d'analyse, la précision des facteurs d'émission peut ne pas être un facteur très important.

Les principaux polluants traditionnellement modélisés par les modèles d'émissions des véhicules sont : le monoxyde de carbone ($CO$), les composés organiques volatils ($COV$), les oxydes d'azote ($NO_x$), le dioxyde de carbone ($CO_2$) et les particules ($PM_{2.5}$ et $PM_{10}$).

Les oxydes d'azote ($NO_x$) ne sont pas des produits directs de la combustion. Cependant, leur production a lieu dans un environnement créé par la combustion. Elle est due à la réaction chimique entre l'azote présent dans l'air atmosphérique et les gaz à haute température formés par la combustion. Les $NO_x$ émis sont composés d'oxyde nitrique ($NO$) et de dioxyde d'azote ($NO_2$), ce dernier étant significativement plus petit en quantité que le premier. Lorsque la consommation de carburant est

faible, une petite quantité de $NO_x$ est émise. L'émission de $NO_x$ dans les moteurs diesel est plus élevée que celle des moteurs à essence. Ceci est dû aux caractéristiques de combustion des moteurs diesel, qui ont des températures et des pressions plus élevées (De Nevers, 2000). $NO_x$ est l'un des précurseurs de la formation d'ozone. De plus, il réagit avec de l'ammoniac et d'autres composants pour former de l'acide nitrique, ce qui peut causer des problèmes respiratoires.

Les modèles d'émissions peuvent être classés comme étant dynamiques ou statiques en fonction de l'échelle de temps utilisée et comme macroscopiques ou microscopiques selon l'échelle spatiale. L'approche du choix de la modélisation dépend de l'objectif et des contraintes de l'étude. Les modèles d'émission peuvent également utiliser des données à partir de modèles de trafic microscopiques. Différentes approches peuvent être utilisées pour alimenter le modèle d'émission avec des vitesses de véhicules et des accélérations. Il est possible, par exemple, d'appliquer une répartition spatiale de la vitesse et/ou de l'accélération en fonction des cycles de conduite ou des distributions statistiques (Burghout, 2004).

## 1.3 Les problèmes de couplage entre les modèles de trafic et d'émission et la précision de l'estimation

En tenant compte des effets de la dynamique du trafic sur le réseau et plus précisément de la précision requise, cela entraîne une augmentation significative des données traitées en volume et du calcul du temps pour obtenir des résultats. Cette complexité est nécessaire lorsqu'il s'agit de décrire une bonne résolution de l'espace et du temps dans l'évolution des émissions. Cela peut sembler excessif, mais quand il est juste de comparer différents projets par rapport à leurs impacts mondiaux.

De nombreux modèles d'émissions existent et sont alimentés par des représentations du trafic spécifique à chaque échelle d'approche (Can and Leclercq, 2009). Seuls les meilleurs modèles sont en mesure de prendre en compte les effets de la dynamique du trafic, *i.e.* phénomènes de congestion et donc évaluer l'impact environnemental de différentes stratégies de contrôle de la circulation. Plus la description des phénomènes de trafic à l'entrée d'un modèle d'émission est importante, plus le volume de calcul à quantifier est important, mais la résolution spatio-temporelle est plus fine. Ce volume de calcul très important peut sembler superflu lorsque l'on s'intéresse seulement à la quantification totale des émissions atmosphériques dans une zone, mais il faut néanmoins tenir compte des effets de la dynamique du trafic et des variations locales des conditions de circulation. La question, cependant, est de savoir si une méthode d'échantillonnage efficace n'atteindrait pas les mêmes résultats en gardant une description précise des phénomènes, mais seulement pour un sous-échantillon. Cette thèse traite spécifiquement de cette question afin d'améliorer les méthodes d'évaluation de l'impact des projets de développement routier ou des stratégies de contrôle de la circulation.

Les erreurs dans l'estimation des émissions des véhicules peuvent conduire à la mise en oeuvre de stratégies de gestion du trafic ou de l'organisation des transports qui ne sont pas nécessairement les plus efficaces en termes de qualité de l'air. Par conséquent, les erreurs liées au couplage entre les modèles d'émission et de trafic doivent donc toujours être liées au gain attendu par la mise en œuvre de stratégies de trafic qui peuvent être inférieures à 4% (Font and Fuler, 2016). La précision de l'estimation des émissions du couplage entre les modèles de trafic et d'émission est affectée par deux types d'erreurs : (i) les erreurs associées à la fiabilité des données de trafic et (ii) les erreurs inhérentes à la modélisation des émissions des véhicules. Ces deux points sont également abordés dans cette thèse.

# 2 Analyse descriptive d'un réseau de transport - le cas de Paris

## 2.1 Introduction

Ce chapitre s'intéresse plus particulièrement aux données simulées où seuls les principaux résultats et conclusions sont abordés. Une partie du 6ème arrondissement de Paris, a servi de base à notre étude. Le réseau a été construit dans le cadre du projet ISpace & Time (Villegas et al., 2013) financé par l'ANR (l'Agence Nationale de Recherche nationale). Le réseau orienté se compose de 234 liens, 93 carrefours, 19 entrées, 21 sorties comprenant 4 parcs de stationnement et 27 feux de signalisation. Ce réseau a été implémenté dans le simulateur de trafic microscopique Symuvia développé par le LICIT (Laboratoire Ingénierie Circulation Transport). Ce simulateur de trafic est utilisé pour définir les paramètres de trafic qui représentent, de la façon la plus réaliste, les conditions de trafic sur un réseau donné. Trois paramètres principaux doivent être pris en compte : l'évolution temporelle de la demande, la matrice origine-destination et la matrice d'affectation. Nous considérons une fenêtre temporelle afin d'éviter un long temps de calcul pour simuler 24 heures de trafic, on considère afin de réduire les temps de calcul, où celle-ci a été fixée aux 6 heures les plus représentatives du trafic quotidien. L'évolution temporelle de la demande est représentée par le trafic de deux heures de pointe : le matin et le soir. Le premier correspond à la demande intense distribuée dans un court laps de temps tandis que le pic du soir a une demande modérée répartie dans un temps plus long.

Les voitures particulières sont l'unique mode retenu dans notre modélisation du trafic. Notons que les informations du trafic sont agrégées par période de 15 minutes.

De plus, deux types de capteurs virtuels sont utilisés pendant la simulation. Les boucles spatiales fournissent une information complète du trafic local comme : le temps de déplacement total, les distances parcourues par les voitures et leur vitesse moyenne. Les boucles inductives donnent quand à eux le nombre de véhicules et la vitesse moyenne observée au milieu de chaque lien. Ce type de boucle est un exemple de données de trafic utilisées par les gestionnaires de ville pour évaluer leurs stratégies de régulation des réseaux. Il est important de noter que les informations de trafic spatial au niveau du lien proviennent uniquement des capteurs spatiaux. Cela ne peut être dérivé que par simulation. Au final, le niveau d'information le plus fin possible en simulation microscopique est celle du lien et de la période. Une analyse de sensibilité entre les deux types de capteurs pour quantifier les émissions sera discutée dans la section suivante.

Pour être statistiquement représentatif, un nombre important d'observations (c'est-à-dire des simulations) et différents états de trafic dans l'espace et le temps sont nécessaires. Pour cela„ le nombre de simulations a été fixé à 400 représentant plus d'un an de données de trafic dans la région étudiée. Notons que chaque simulation est caractérisée par une demande stochastique pour chaque entrée et période de temps.

L'état de l'art des modèles d'estimation d'émissions est dense, bal ayant plusieurs échelles spatiales et temporelles, et plusieurs niveaux de granularité. L'échelle microscopique est le niveau le plus fin, avec une représentation particulière (comme un véhicule ou une rue) et l'échelle macroscopique est le niveau le plus agrégé (comme une région ou une nation). Dans la littérature, les paramètres d'entré les plus usuels sont : les polluants "couverts", les type d'émissions, la composition de la flotte (les

catégories de véhicules et l'âge), les motifs de conduite (la vitesse moyenne ou les vitesses instantanées et leurs accélérations). Dans notre cas d'étude, nous utilisons le modèle d'émission COPERT IV basé sur l'intégralité des variables précédemment énumérées.

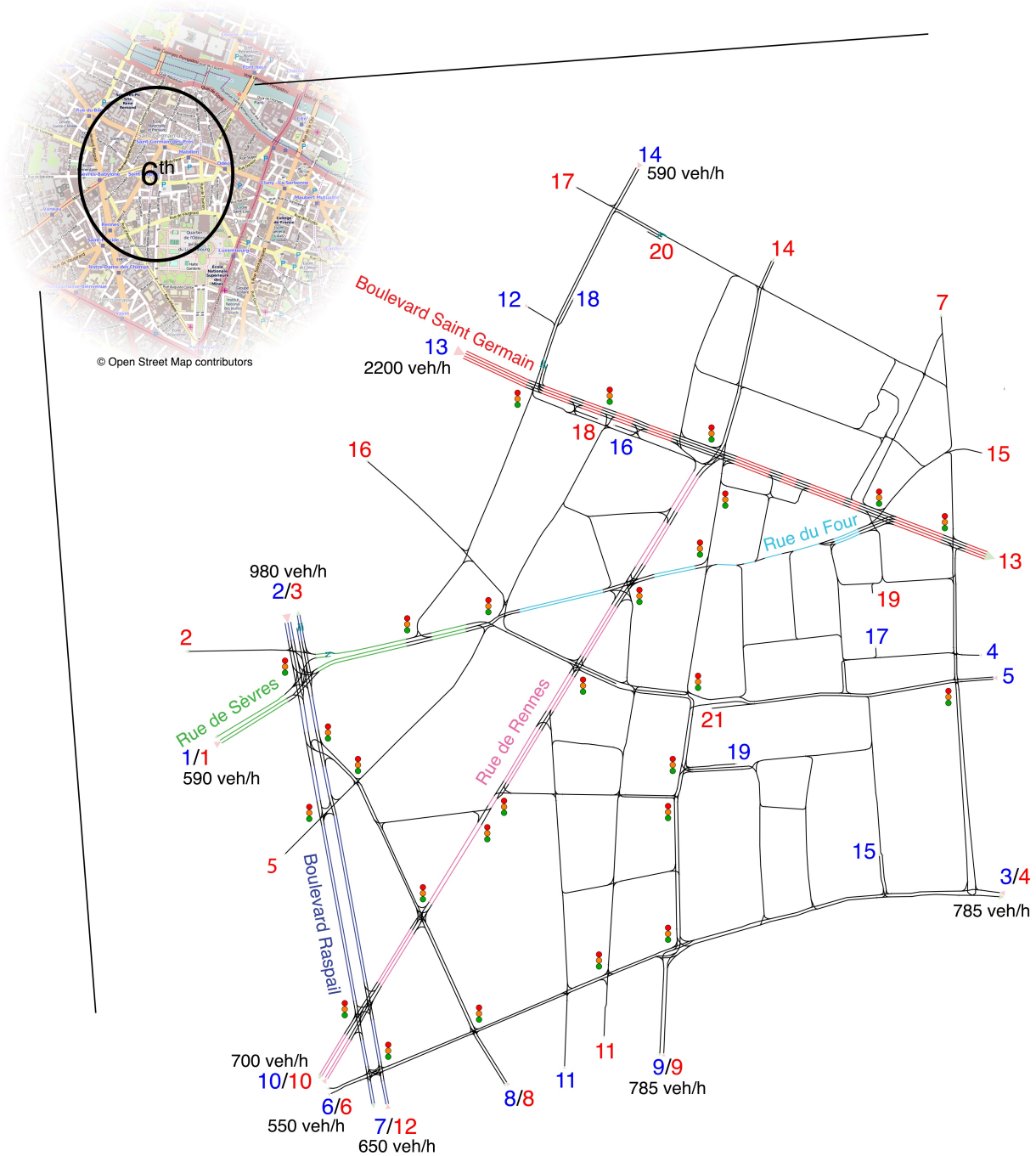



Fig. 2.1 – *Réseau de transport routier du 6éme arrondissement de Paris. Les principaux axes du réseau sont indiqués avec des couleurs. Les valeurs de la demande aux heures de pointe sont affichées dans à chaque entrée.(Villegas et al., 2013)*

## 2.2 Influence de la définition de la variable pour l'estimation des émissions

A partir des donées simulées par SymuVia, les informations du trafic ont été extraites. Plus précisément, les boucles spatiales permettent d'obtenir 3 informations du trafic (pour chaque lien et chaque période) : le temps de déplacement total, la distance totale parcourue et la vitesse moyenne. Le temps de déplacement total et la distance totale parcourue (c'est-à-dire la production de voyage) désignent respectivement, le temps total passé par les véhicules et la distance totale parcourue à travers le réseau de transport pendant la période réglée. Les vitesses moyennes spatiales sont calculées par le rapport de la distance totale parcourue et du temps total passé par les véhicules pour chaque lien et à chaque période.

Les boucles inductives donnent quand à eux des flux de véhicules et des vitesses moyennes à l'emplacement des capteurs (c'est-à-dire dans notre cas au milieu de chaque voie de chaque tronçon du réseau). Pour calculer les émissions à l'aide de COPERT IV, deux informations du trafic sont requises : la distance parcourue et la vitesse moyenne. En utilisant les boucles magnétiques comme source d'information du trafic, les distances parcourues doivent être calculées à partir des flux de véhicule. Pour cela, deux définitions de longueur de lien seront utilisées : (i) une approche statique que considère la longueur du tronçon géométrique. Il s'agit de la longueur entre le début et la fin du tronçon (incluant la distance entre la sortie du lien et le barycentre du carrefour en amont, si elle a lieu). (ii) une approche dynamique considère les distances supplémentaires des mouvements autorisés à l'intérieur du carrefour. Ce dernier permet de connaître la distance réelle parcourue par les véhicules sur le lien et à l'intérieur du carrefour, selon le débit, au lieu de les estimer en utilisant des mesures géométriques. Il est intéressant de comprendre que la longueur de tronçon géométrique est une grandeur statique et ne dépend pas du flux de trafic. Contrairement à cela, la longueur de tronçon dynamique dépend complètement du flux de trafic sur chaque lien et carrefour.

Nous considérons deux polluants, le $CO_2$ (dioxyde de carbone) qui a le plus d'impact sur l'effet de serre et les $NO_x$ (oxydes d'azote) qui ont une incidence sur la santé publique. L'évaluation des émissions se fait selon le choix des paramètres tels que la composition de la flotte, le type d'émissions et les émissions dépendant de la vitesse. La composition de la flotte française de 2015 a été choisie et l'étude se concentrera sur les émissions chaudes. Pour calculer la quantité de chaque polluant, les courbes dépendantes de la vitesse seront utilisées. Ces dernières fournissent des facteurs d'émission pour chaque vitesse moyenne supérieure à 10 km/h. Considérant que, pour des vitesses moyennes inférieures à 10 km/h, les émissions seront calculées en utilisant un facteur d'émission égal à 10 km/h. L'équation qui sera utilisée pour quantifier les émissions avec la courbe de vitesse est la suivante :

$$e_p = d_i \times Fe_p(\bar{v_i}) \tag{2.1}$$

où :

— $e_p$ est l'émission de polluants en g/km ;

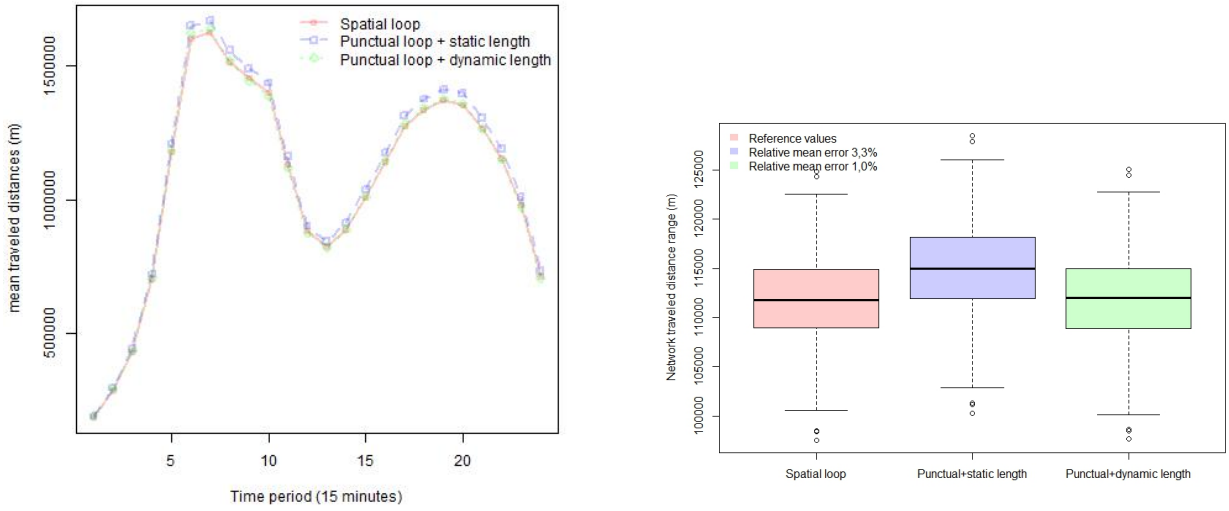
— $d_i$ est la distance parcourue dans l'élément spatial $i$ (lien/tronçon) ;

— $Fe_p$ est le facteur d'émission de polluants associé à un parc automobile déterminé par une vitesse moyenne dans $i$ ;

— $\bar{v_i}$ est la vitesse moyenne en $i$.

La production de déplacement provenant du capteur spatial est utilisée comme référence pour évaluer les méthodes de calcul de la distance totale parcourue à l'aide de données de trafic à partir de boucles inductives (vitesse et débit moyen par lien par période). La figure 2.2 montre la comparaison de ces deux hypothèses avec la référence.

Les données issues de la "Boucle spatiale" sont les valeurs de référence, la "Boucle inductive + longueur statique" et la "Boucle inductive + longueur dynamique" correspondent respectivement aux distances totales parcoures par la approche statique et dynamique. La dernière est la méthode qui correspond le mieux aux valeurs de référence avec seulement 1% d'erreur moyenne (pour les 400 valeurs simulées). Il est intéressant de noter que les distances parcourues par l'approche statique ont presque la même répartition que ceux par l'approche dynamique. En effet, ce dernier n'est pas surprenant car




(a) Distance moyenne parcorue journalières par période de temps.

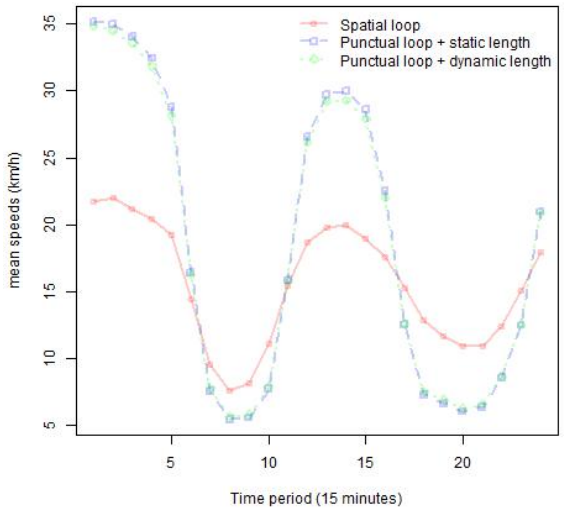
(b) Distribuitions des erreurs de la distance parcorue moyennes journalières et l'erreur relative.

Fig. 2.2 – *Comparaison de distance parcourue dans le réseau.*

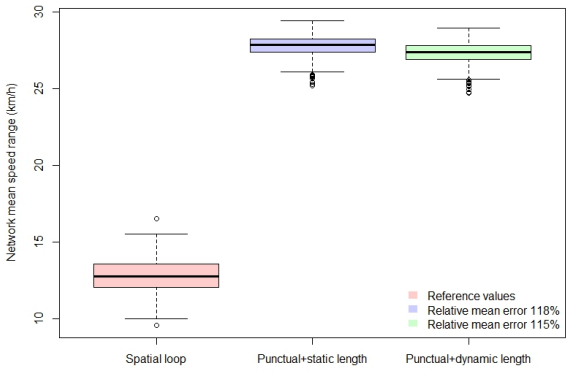
les distances totales parcourues sont le produit entre le nombre de véhicules qui passent au capteur dans un lien donné et la distance parcourue moyenne du même lien. Ainsi, nous supposons que tous les véhicules traversant le capteur ont parcourue la même distance. Par conséquent, cette distance parcourue sera un peu surestimée, à environ 3%, comme le montre la figure 2.2, car elle considère que tous les véhicules traversant le capteur parcourent la totalité de la longueur du tronçon géométrique, alors que parfois, ce n'est pas le cas . Les différences entre elles sont faibles étant donné qu'elles ont été calculées au niveau du réseau (c'est-à-dire dans un temps et un espace agrégés). Cette différence se trouve également dans la figure 2.2 qui montre les valeurs de distribution de chacune et l'erreur moyenne relative de chaque méthode par rapport aux valeurs de référence. Dans une perspective de gestionnaire et compte tenu des faibles erreurs de distances parcourues (moins de 3,5% en moyenne sur 400 jours), la méthode utilisant l'approche statique permet, de manière simple, de déterminer la distance parcourue d'un lien ou d'un réseau directement en utilisant les données collectées par un capteur et des cartes géoréférencées sans avoir à utiliser des simulations à cette fin.

La deuxième variable de trafic qui doit être analysée pour estimer les émissions est la vitesse moyenne. Le réseau à l'étude représente une zone urbaine à faible vitesse moyenne sur 15 minutes et sa variation est comprise entre 1 km/h et 50 km/h localement. Les vitesses sont le rapport entre les distances et les temps et, compte tenu de la faible différence entre la distance parcourue statique (c'est-à-dire l'utilisation de la longueur de tronçon géométrique) et dynamique (c'est-à-dire en utilisant la longueur de tronçon dynamique), les deux donnent presque les mêmes résultats. Les vitesses moyennes des boucles inductives sont à la fois surestimées et ont atteint de grandes erreurs relative, à environ 115% d'erreur moyenne, comme le montre la figure 2.3. A l'échelle du réseau, l'écart des vitesses moyennes mesurées entre les boucles spatiales et inductives sont comparées. La gamme des vitesses moyennes varie entre 5 et moins de 35 km/h. Notons que ces faibles vitesses pratiquées sont totalement normales pour la zone urbaine représentée. En outre, ces faibles vitesses ont une importance lorsque les émissions sont calculées, car elles ont des facteurs d'émission plus élevés. Les grandes différences entre la vitesse moyenne des boucles spatiales et des boucles inductives sont plus évidentes avec des états de trafic fluide (sur influence des feux de circulation) et peuvent atteindre 14 km/h de différence. Cette différence varie en fonction de l'état de la circulation. Ce fait explique notamment pourquoi la vitesse moyenne est considérée au niveau du lien. Pour les boucles inductives, la vitesse moyenne considérée pour toutes les longueurs de tronçon est mesurée à partir d'un point au milieu de chaque voie. Comme la plupart des liens ont une petite longueur, les véhicules qui traversent le capteur s'accélèrent encore. Contrairement aux boucles inductives, les boucles spatiales calculent la vitesse moyenne en tenant

compte de la longueur totale de chaque tronçon (approche spatiale) et non d'un point et cela n'est possible que par des simulations.




(a) Vitesses moyennes du réseau par période de temps.




(b) Distribuition de l'erreur par vitesse moyenne du réseau.

Fig. 2.3 – *Comparaison des vitesse moyennes du réseau.*

A l'échelle du réseau, nous utilisons un indicateur agrégé étant la somme des mesures d'émissions par lien et par période de 15 minutes.

Cette méthode a été utilisée pour obtenir les émissions de capteurs spatiaux et des capteurs inductifs, sachant que la dernière a deux options de productions de voyage, ce qui donne en conséquence, deux possibilités de valeurs d'émission.

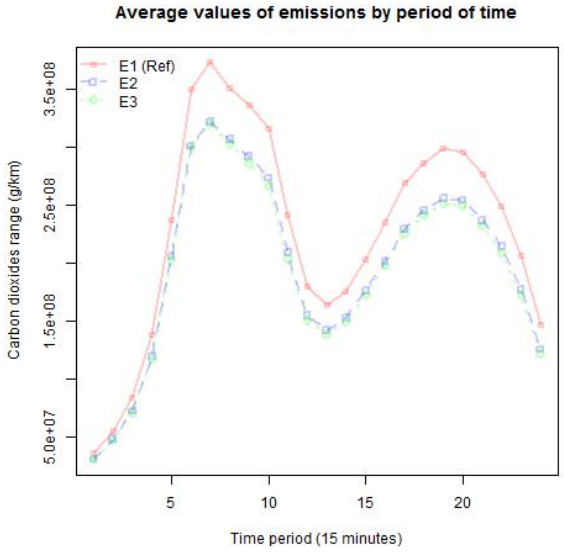
La figure 2.4 compare les émissions de polluants des deux capteurs : (a) et (b) correspondent aux émissions du réseau de dioxyde de carbone ; et (c) et (d) pour les émissions de $NO_x$. Comme on peut le voir, les émissions de polluants calculées à l'aide des données de trafic locales provenant des capteurs inductifs présentent des valeurs inférieures à celles des capteurs spatiaux. Ces faibles quantités d'émissions sont dues au fait que les capteurs inductifs considèrent des vitesses beaucoup plus élevées que le capteur spatiaux et par conséquent des facteurs d'émission plus faibles ; ces différences sont plus évidentes en état de congestion.

Comme le montre la figure 2.3, les vitesses moyennes du réseau à partir des capteurs inductifs se situent entre 5 et 35 km/h, au lieu de 7 et 25 km/h pour les capteurs spatiaux, par conséquent, les valeurs à grande vitesse ont tendance à avoir un coefficient d'émission plus faible.

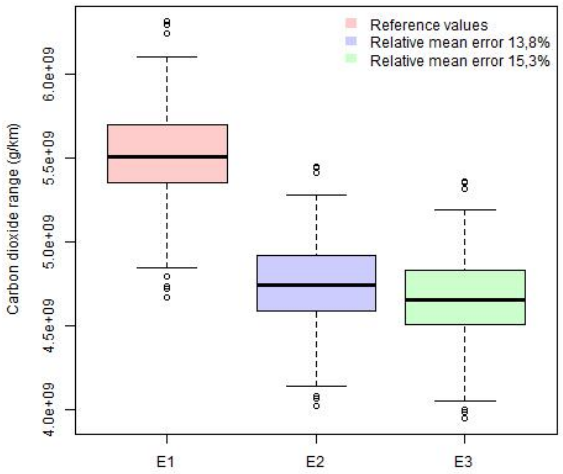
L'émission de polluants est le produit de la distance parcourue et du facteur d'émission correspondant pour un polluant donné déterminé par la vitesse moyenne. La différence entre les trois distances parcourues est très faible (figure 2.2), mais la comparaison de la moyenne de la vitesse montre des vitesses différentes dans les deux capteurs et qui se termine par une sous-estimation d'environ 14% des émissions du réseau en utilisant les données de trafic de ce capteur ((b) et (d) dans la figure 2.4).

## 2.3 Conclusions

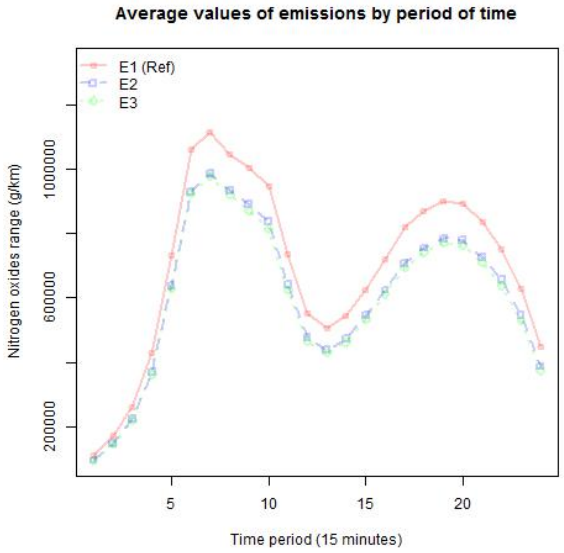
Dans ce chapitre, une analyse descriptive du cas d'étude a été réalisé basée sur des données simulées de plus d'un an d'information sur le trafic. Le réseau et l'environnement de simulation ont été décrits. L'émission de pollunt par le trafic routier est la variable d'analyse où deux types de mesures sont considérés : les capteurs spatiaux et inductifs. Une comparaison a été faite pour identifier la meilleure description de variable à utiliser comme base du travail pour les chapitres suivants. Cette étude a montré que les informations de trafic à partir de capteurs inductifs peuvent fournir des valeurs biaisées.



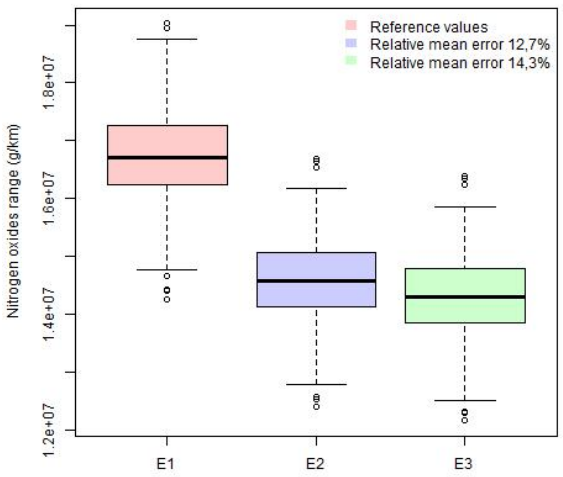

(a) Les emissions de $CO_2$ du réseau par période de temps.




(b) Distribuition de l'erreur des emissions de $CO_2$ du réseau.




(c) Les émissions de $NO_x$ du réseau par période de temps.




(d) Distribuition de l'erreur des émissions de $NO_x$ du réseau.

Fig. 2.4 – *Comparaison des émissions des polluants du réseau.*

Les valeurs de la distance totale parcourue sont surestimées de 3% en moyenne et ce biais est accentué en périodes congestionnées. Compte tenu de la vitesse moyenne, la surestimation peut atteindre plus de 100% d'erreur. Pour des états de trafic fluide, cette différence peut être augmentée et atteindre 14 km/h. Les deux biais de trafic peuvent induire une sous-estimation des émissions du réseau en moyenne de 14%.

Les informations routières provenant de capteurs inductifs sont généralement utilisées par les gestionnaires de la ville pour appliquer de nouvelles stratégies ou pour mener des études sur la mobilité et les réductions d'émissions. Cependant, la précision et la fiabilité des informations de trafic fournies par la simulation microscopique sont nécessaires pour estimer avec précision les émissions des polluants et pour étudier leur variabilité selon les échelles spatiales et temporelles. Ainsi, les données spatiales ont été retenues pour étudier et appliquer les méthodes d'échantillonnage proposées dans le cadre de cette thèse.

# 3

# Échantillonage des liens par la méthode LASSO

## 3.1 La méthode LASSO

Le LASSO est une méthode de régression qui implique de pénaliser la taille absolue des coefficients de régression. En pénalisant (ou en contraignant de manière équivalente la somme des valeurs absolues des estimations) certaines estimations de paramètres peuvent être exactement nulles. Plus la pénalité est importante plus les estimations supplémentaires sont réduites vers zéro.

Least absolute shrinkage and selection operator (LASSO) est une méthode statistique moderne qui a attiré beaucoup d'attention au cours de la dernière décennie car les chercheurs dans de nombreux domaines sont capables de mesurer beaucoup plus de variables qu'auparavant. La régression linéaire comporte un inconvénient majeur : le nombre de prédicteurs devient grand. Dans ce cas, non seulement un dépassement peut se produire, ce qui signifie que le modèle adapté ne se généralise pas efficacement au-delà des données particulières observées ; mais, il devient aussidifficile d'interpréter les modèles adaptés. LASSO aborde ces deux problèmes en identifiant un petit nombre de prédicteurs sur lesquels un modèle fiable peut être construit.

Nous avons un ensemble de variables explicatives $(X_i)$ pour $1 < n < p$ pour expliquer une variable $Y$ linéairement, et rien ne garantit que toutes les variables impliquées sont explicatives. Nous avons donc un ensemble de variables ou de candidats potentiellement explicatifs. Notre but est d'identifier les variables explicatives. Par conséquent, il est nécessaire de choisir un modèle parmi les possibilités de $2^p$. Comment choisir le bon modèle ? Il faut étudier toutes les possibilités qu'il ne sont pas possibles lorsque $p$ est grand, et plus important encore, savoir quel modèle est meilleur que les autres. La méthode LASSO offre, dans certains cas, une solution à ce problème. Ceci est pratique lorsqu'il s'agit de prédicteurs hautement corrélés, où la régression standard aura habituellement des coefficients de régression «trop importants» (Tibshirani, 1996).

Nous cherchons à expliquer une variable $Y$ linéairement par des variables $p$ potentiellement explicatives de $X_i$. À cette fin, nous simulons $n$ observations. Le modèle variable $Y$ est décrit ci-dessous :

$$Y = X\beta + \varepsilon \tag{3.1}$$

où $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ est un vecteur de variables aléatoires de $n$ avec une valeur moyenne égale à zéro et une variance $\sigma^2$ qui correspond au bruit dans les observations (*c.-à-d.* qui peuvent contenir toutes les variables explicatives non prises en compte dans le modèle) ; $Y \in \mathbb{R}$ est un vecteur qui correspond à $n$ observations de $Y$ (*c.-à-d.* les valeurs de réponse que nous voulons prédire). $X = (X_{.,1}, \cdots, X_{.,p}) = ((X_{1,.})^T, \cdot, (X_{n,.})^T)^T$ est une matrice $n \times p$, où $n$ sont les lignes de la matrice et correspondent aux valeurs d'observation du prédicteur $X_p$ ; $p$ est la colonne de la matrice qui correspond aux variables $X$. Mathématiquement parlant, $X_p$ est la p$^{\text{ième}}$ colonne qui correspond au p$^{\text{ième}}$ prédicteur de $X_p$. $X_n$ est la n$^{\text{ième}}$ ligne qui correspond à la n$^{\text{ième}}$ observation. $\beta \in \mathbb{R}^p$ est le paramètre qui doit être estimé et indexé par $n$ pour permettre à ces coefficients et sa taille de varier quand $n$ augmente ($p$ peut dépendre de $n$).

Si les variables $X_p$ ne sont pas toutes pertinentes, l'objectif est d'éliminer les variables inutiles et seulement celles-ci. L'idée de LASSO n'est pas d'effectuer une régression linéaire classique, mais une

régression régulière qui fait que certains des coefficients $\beta$ sont égaux à zéro. Cela implique l'estimation $\lambda \in \overline{\mathbb{R}}_+$. Compte tenu de ce dernier, les coefficients $\beta$ sont calculés comme suit en considérant $\lambda \in \overline{\mathbb{R}}_+$.

$$\widehat{\beta}(\lambda) = \underbrace{argmin}_{\beta \in \mathbb{R}^p}(\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1) \tag{3.2}$$

où $\|x\|_2^2 = \sum_{i=1}^n x_i^2$ et $\|x\|_1 = \sum_{i=1}^p |x_i|$.

Le paramètre $\lambda \geq 0$ contrôle la puissance de la régularisation. Si $\lambda = 0$, la méthode LASSO correspond à une régression linéaire classique (si $p \geq n$). Au contraire, si $\lambda = \infty$, tous $\widehat{\beta}(\infty)$ sont égaux à zéro. L'augmentation de $\lambda$ induit certains coefficients $\widehat{\beta}(\lambda)$ pour diminuer jusqu'à zéro. Le dernier modèle équivaut à ce qui suit :

$$\widetilde{\beta}(t) = \underbrace{argmin}_{\beta, \|\beta\|_1 \leq t}(\|Y - X\beta\|_2^2) \tag{3.3}$$

considérant pour tout $\lambda \in \overline{\mathbb{R}}_+$, $t \geq 0$ tel que : $\widetilde{\beta}(t) = \widehat{\beta}(\lambda)$. En effet, il suffit de prendre $t = \|\widehat{\beta}(\lambda)\|_1$ puis pour tout $\beta$ tel que $\|\beta\|_1 \leq t$, $\lambda\|\beta\|_1 \leq \lambda\|\widehat{\beta}(\lambda)\|_1$ donc, en définissant $\widehat{\beta}(\lambda)$, $\|Y - X\beta\|_2^2 \geq \|Y - X\widehat{\beta}(\lambda)\|_2^2$.

Cette explication nous permet de comprendre intuitivement pourquoi, dans la plupart des cas, LASSO aboutit exactement à zéro pour certains coefficients $\widehat{\beta}(\lambda)$ (Friedman et al., 2010).

L'algorithme qui a été appliqué pour résoudre LASSO utilise la descente de coordonnées cycliques calculée le long du chemin de régularisation (Friedman et al., 2010). Il consiste à déterminer $\widehat{\beta}(\lambda)$ pour tout $\lambda \geq 0$. La prochaine étape consiste à déterminer le $\lambda$ qui ne peut contenir que des variables explicatives réelles et à éliminer les autres. Une approche générale consiste à utiliser une erreur de prédiction pour guider ce choix. L'une de ces méthodes s'appelle la validation croisée.

La validation croisée fonctionne en divisant les données de formation au hasard en dix parties égales. La méthode d'apprentissage est adaptée à une gamme de valeurs de $\lambda$ (c'est-à-dire paramètre de complexité) - à neuf dixièmes des données, et l'erreur de prédiction est calculée sur le dixième restant. Ceci se fait à tour de rôle pour chaque dixième des données, et les dix estimations d'erreur de prédiction sont calculées en moyenne. À partir de cela, nous obtenons une courbe d'erreur de prédiction estimée en fonction du paramètre de complexité. Il est toujours nécessaire de diviser les données en un jeu d'entrainement et un jeu de validation. La validation croisée est appliquée à l'ensemble d'entrainement, car la sélection du paramètre de retrait fait partie du processus d'entrainement. L'ensemble de validation est là pour juger de la performance du modèle sélectionné.

L'algorithme choisi calcule tout un chemin de solutions dans $\lambda$ pour n'importe quel modèle particulier, laissant l'utilisateur sélectionner une solution particulière de l'ensemble. Il est possible d'évaluer la performance de prédiction à chaque valeur de $\lambda$ et de choisir le modèle avec la meilleure performance. Pour évaluer les modèles, l'erreur de prédiction moyen-carré a été utilisée comme mesure du risque. À partir de la courbe d'erreur moyenne obtenue grâce à la validation croisée, deux modèles sont mis en surbrillance, une ligne qui correspond à l'erreur minimale et donne un modèle avec le nombre $p$ de prédicteurs ,et une autre ligne qui donne la plus grande valeur de lambda de telle sorte que l'erreur est dans une seule norme -direction de l'erreur minimale - la soi-disant "règle d'erreur standard". Les deux modèles en surbrillance ont les meilleures performances, mais la différence entre eux est le nombre de prédicteurs sélectionnés sur chaque modèle. La règle d'erreur standard unique donne un modèle avec moins de prédicteurs que le modèle défini par une erreur minimale. Lorsque nous comparons les valeurs d'erreur des parents des valeurs prédites sur les deux modèles, elles ont une erreur similaire. Pour cette raison, le modèle utilisant la règle «une erreur standard» est le meilleur choix car il y a un nombre minimal de régresseurs sélectionnés et, à partir de là, toutes les analyses seront effectuées. La comparaison entre les deux modèles pour confirmer le choix fait est détaillée dans (Friedman et al., 2010).

## 3.2 Les jeux de données

Trois types d'ensembles de données ont été construits pour aider à caractériser le comportement dynamique du réseau, dans ce résumé ne seront présentés que les résultats d'un des jeux de données. Ils ont été construits pour chaque variable, *c.-à-d.* la distance parcourue totale, la vitesse moyenne, les émissions de $CO_2$ et de $NO_x$. Les structures du jeu de données présentes dans ce résumé sont expliquées ci-dessous.

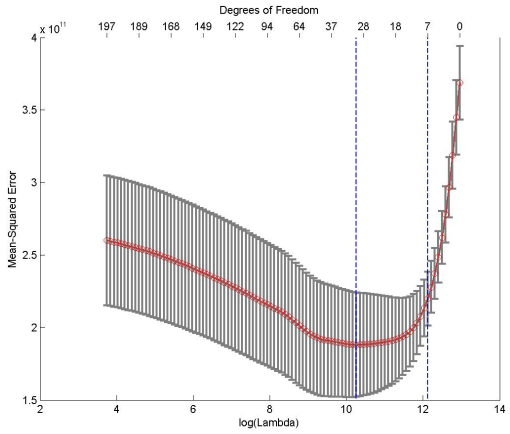
L'ensemble de données appelé statique ne considère que les valeurs quotidiennes de trafic pour chaque lien dans le réseau, *c.-à-d.* la distance parcourue quotidiennement, la vitesse moyenne et les émissions.

Chaque valeur d'observation de chaque lien du réseau a été fournie par une simulation. Dans le modèle, les régresseurs sont les liens et leurs observations sont la distance totale parcourue, la vitesse moyenne et les valeurs totales d'émission ($CO_2$ et $NO_x$) pour chaque simulation (qui représentent les 6 heures les plus pertinentes du jour).

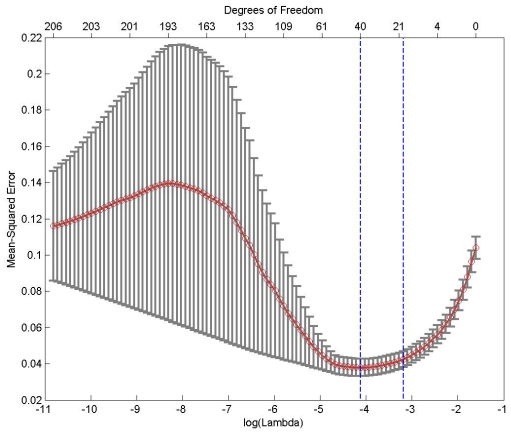
L'objectif de cet ensemble de données est d'estimer au niveau du réseau la quantité d'émissions par jour en utilisant uniquement le trafic quotidien et les émissions provenant de quelques liens du réseau. Cette méthode est proposée comme une simplification du processus pour évaluer les émissions dans le réseau en utilisant uniquement les valeurs quotidiennes du trafic réseau et des émissions à partir de l'ensemble des liens sélectionnés.

## 3.3 Résultats

Comme expliqué précédemment, le modèle proposé par $\lambda$ a un écart type de l'erreur carrée minimale était le modèle retenu pour toutes les variables. Les résultats ne sont présentés que pour ce modèle. Dans la figure 3.1, on montre le résultat de la validation croisée de $\lambda$ validé en croix pour chaque variable dans l'ensemble de données statiques. La figure 3.1 représente les courbes d'erreur de prédiction estimées et leurs écarts types pour les variables dans les ensembles de données statiques : $(a)$ le modèle est réglé pour la distance parcourue, $(b)$ pour la vitesse moyenne, $(c)$ pour le $CO_2$ et $(d)$ pour les émissions de $NO_x$ . Chaque courbe est tracée en fonction du paramètre de complexité correspondant $\lambda$. L'axe horizontal a été choisi de sorte que la complexité du modèle augmente à mesure que nous passons de droite à gauche. Les estimations de l'erreur de prédiction et leurs erreurs-types ont été obtenues par une validation croisée de dix fois. Le modèle le moins complexe (le plus petit modèle parmi les modèles optimaux), indiqué par les lignes verticales. La ligne verticale gauche est le modèle avec une erreur minimale et la droite est le modèle réglé en utilisant la règle de l'écart type (le modèle qui sera étudié). Le haut de chaque figure est annoté avec la taille des modèles.

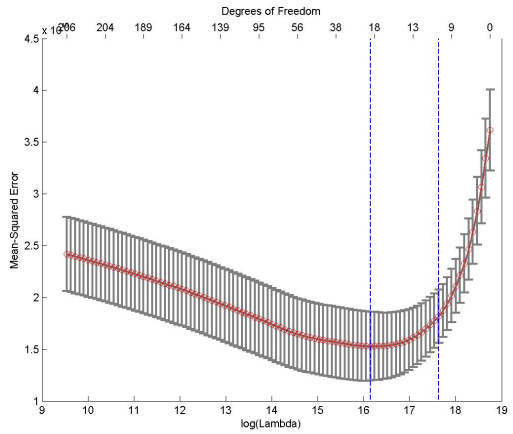



(a) Courbe d'erreur de prédiction moyen-carré de la distance parcourue.

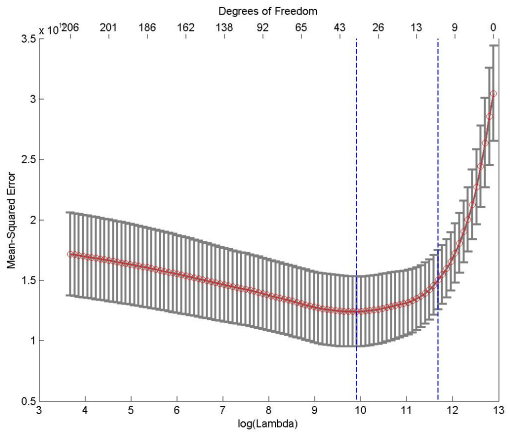



(b) Courbe d'erreur de prédiction moyen-carré de vitesse moyenne spatiale.

Le LASSO a fait une sélection sur plus de 230 liens du réseau et nous donne les modèles suivants : pour a distance parcourue, le modèle comporte 7 liens et peut expliquer 43% des données compte tenu




(c) Courbe d'erreur de prédiction moyen-carré de l'émission de $CO_2$.




(d) Courbe d'erreur de prédiction moyen-carré de l'émission de $NO_x$.

Fig. 3.1 – *Estimation des courbes d'erreur de prédiction pour les variables dans le jeu de données statiques/statiques.*

de l'intervalle de confiance à 95% ; "à la ligne" pour la vitesse moyenne, le LASSO a sélectionné 19 liens avec 64% des données expliquées par le modèle ; "à la ligne" pour les émissions de polluants, un modèle comprenant 11 liens explique 54% des données dans les émissions de CO2 et 55% des données dans les émissions de $NO_x$. Les erreurs relatives ont été calculées en comparant les résultats (valeurs prédites par le modèle établi) avec les valeurs de référence ($Y$). La figure 3.2 montre la répartition des erreurs associées dans chaque variable pour l'ensemble de données statiques.

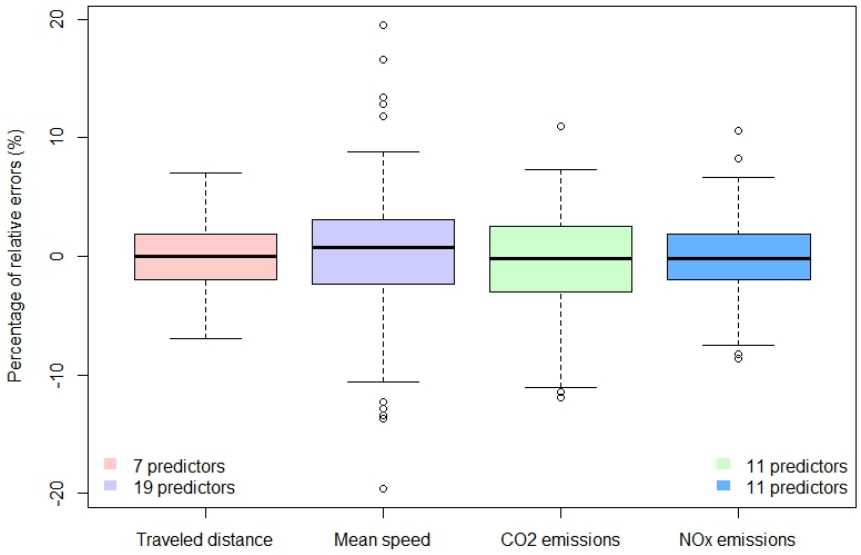



Fig. 3.2 – *Pourcentage de répartition des erreurs entre les valeurs des variables prédites par le modèle construit avec des liens sélectionnés et les valeurs de référence des variables (Y ).*

Toutes les variables ont de petites erreurs moyennes lorsque l'on considère des modèles qui ont moins de 9% des liens du réseau sélectionné. Plus de 50% des données ont des erreurs inférieures à ±5% et, en considérant toutes les données, les erreurs atteignent un peu plus de ±10% au plus pour toutes les variables.

Une analyse croisée a été menée pour observer si l'un des 4 modèles pourrait être utilisé pour

déterminer également autres variables. L'objectif est d'avoir un ensemble de liens sélectionnés qui peut être utilisé pour quantifier les valeurs du réseau pour toutes les variables. Le tableau ci-dessous montre le pourcentage moyen d'erreur du modèle de liens sélectionné établi par les variables dans les colonnes appliquées sur les variables disposées dans les lignes.

| Daily average percentage of error (absolute error) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP | 7 | 3,04% | **2,70%** | **2,47%** | 6,29% | 5,87% | 2,39% | 2,73% | 2,40% | 1,92% |
| VIT | 19 | 8,26% | 2,08% | 1,97% | **4,12%** | **4,05%** | 2,36% | 2,69% | 2,32% | 2,22% |
| $CO_2$ | 11 | 4,78% | 1,95% | 1,63% | 6,21% | 5,85% | **3,01%** | **3,41%** | 2,31% | 1,93% |
| $NO_x$ | 11 | 4,78% | 1,93% | 1,65% | 5,38% | 5,26% | 2,23% | 2,82% | **2,87%** | **2,55%** |
| Number of observations | | | *267* | *133* | *267* | *133* | *267* | *133* | *267* | *133* |

Tab. 3.1 – *L'erreur absolue moyenne du modèle établie pour une variable et appliquée à une autre.*

Le tableau 3.1 montre l'erreur moyenne sur le jeu d'entrainement et de validation obtenues avec des liens sélectionnés pour une variable et des coefficients ajustés à une autre variable. Le but ici est d'étudier la possibilité d'utiliser un ensemble de liens sélectionnés pour déterminer toutes les autres variables. À cette fin, dans le même jeu de donées, une régression linéaire a été effectuée sur les liens sélectionnés par LASSO. L'objectif est de trouver les valeurs bêta de chaque lien sélectionné adapté à la variable étudiée. En général, pour tous les cas, les valeurs moyennes des erreurs restent dans la même plage que la méthode LASSO. Les liens sélectionnés en commun pour les 4 variables ont également été comparés et le ratio est présenté dans le tableau 3.2 pour compléter l'analyse.

| VARIABLES → | DTP | VIT | $CO_2$ | $NO_x$ |
|---|---|---|---|---|
| MODELS ↓ | Validation set | Validation set | Validation set | Validation set |
| DTP | **100%** | 0% | 57,1% | 42,8% |
| VIT | 0% | **100%** | 0% | 0% |
| $CO_2$ | 36,4% | 0% | **100%** | 72,7% |
| $NO_x$ | 27,3% | 0% | 72,7% | **100%** |

Tab. 3.2 – *Ratio des liens communs sélectionnés entre les variables.*

Il est possible d'observer dans le tableau 3.2 que la distance parcourue et les variables de vitesse moyenne spatiale n'ont pas de liens sélectionnés communs. Cela s'explique par leur comportement opposé : la distance parcourue est une variable linéaire sur les liens tandis que la vitesse moyenne spatiale ne l'est pas. À la lumière de ces considérations, deux conclusions peuvent être observées : (i) la forte corrélation entre la distance parcourue et la vitesse moyenne spatiale permet de déterminer chaque émission à partir de leur échantillonnage, car elles dépendent de ces deux variables de trafic ; et (ii) le fait que les deux peuvent être utilisées pour déterminer d'autres valeurs de variables à l'aide d'une régression linéaire simple, nous amène à conclure qu'il n'existe pas qu'un seul échantillonnage acceptable (ensemble de liens). Ainsi, il existe de nombreuses preuves de la flexibilité de la sélection. Le modèle avec le moins de liens sélectionnés sera le meilleur choix, en particulier du point de vue pratique, pour les gestionnaires de transport lorsqu'ils décident d'équiper des liens sur le réseau.

Tous les modèles définis par LASSO ou par régression linéaire ont été validés sur un ensemble de validation complètement différent de l'ensemble de données d'entrainement utilisé pour les appliquer. Pourtant, pour pouvoir comparer les résultats, les données d'entrainement et les données de validation sont les mêmes tout au long de cette étude. Les erreurs moyennes restent dans la même gamme avec

les quatre variables : moins de 7% d'erreur. Ainsi, divers liens échantillonnés pourraient fournir une estimation des variables de trafic et d'émission sur le réseau avec une erreur raisonnable.

Compte tenu du faible taux d'échantillonnage dans chaque variable, ainsi que de leurs faibles erreurs de moyenne et en tenant compte du fait qu'elles possèdent des liens communs sélectionnés, on a également envisagé d'étudier la possibilité de créer l'union et/ou l'intersection entre les variables de trafic et entre les polluants des émissions. Par exemple, les liens sélectionnés identifiés par la méthode LASSO pour les données de trafic, la distance totale parcourue et la vitesse moyenne spatiale seront regroupés (union des liens sélectionnés entre deux variables) pour appliquer une régression linéaire et obtenir un nouveau modèle avec les valeurs du coefficient $\beta_i$ ajusté pour chaque prédicteur (liens). De la même manière, l'union entre $CO_2$ et $NO_x$ a été considérée. Compte tenu du fait que certaines variables ont des liens communs, elle a été considérée comme l'intersection entre elles. L'avantage de l'intersection entre eux est la possibilité d'avoir un modèle avec moins de prédicteurs que le modèle établi par l'union d'entre eux. Les erreurs associées ont été quantifiées pour chaque valeur de variable résultante. L'erreur moyenne de chaque régression linéaire est indiquée dans le tableau 3.3 et a été calculée pour chaque variable compte tenu de chaque situation (union ou intersection).

| Daily average percentage of error (absolute error) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARIABLES → | Model size | | DTP | | VIT | | $CO_2$ | | $NO_x$ | |
| MODELS ↓ | Number of links | Network % | Training set | Validation set | Training set | Validation set | Training set | Validation set | Training set | Validation set |
| DTP∪VIT | 26 | 11,30% | 1,92% | 1,78% | 3,01% | 3,48% | 2,24% | 2,55% | 2,23% | 2,18% |
| DTP∩VIT | - | - | - | - | - | - | - | - | - | - |
| $CO_2 \cup NO_x$ | 14 | 6,09% | 1,91% | 1,62% | 5,28% | 5,25% | 2,21% | 2,80% | 2,21% | 2,12% |
| $CO_2 \cap NO_x$ | 8 | 3,48% | 2,00% | 1,68% | 7,00% | 6,61% | 2,41% | 2,98% | 2,38% | 2,10% |
| Number of observations | | | *267* | *133* | *267* | *133* | *267* | *133* | *267* | *133* |

Tab. 3.3 – *Le pourcentage moyen d'erreur absolue du modèle de régression linéaire adapté à la fusion ou à l'intersection entre des variables de même nature.*

Les liens sélectionnés pour les variables "vitesse moyenne" et "distance totale parcourue" sont complètement différents, donc ils n'ont pas de liens communs. L'union des variables de trafic permet d'estimer toutes les variables avec la même précision que la sélection LASSO appliquée séparément. La répartition des erreurs varie de 0,1% à moins de 7% dans l'intervalle de confiance à 95% en général pour toutes les variables. Lorsque la variable distance parcourue (dans le tableau 3.1) est comparée à la sélection par l'union entre les variables de trafic et aussi à la sélection par l'union des émissions de polluants, il est possible d'observer que la répartition des erreurs est moins dispersée avec l'union des liens sélectionnés. Au contraire, les modèles d'union ont plus de liens que le modèle établi par la variable "distance parcourue", ce qui explique que les erreurs soient moins dispersées.

Lorsque la même comparaison est faite avec l'intersection entre les liens sélectionnés par les variables de polluants, ils ont presque la même quantité de liens sélectionnés et des valeurs moyennes d'erreur similaires. Si l'on compare les modèles des polluants dans le tableau 3.1 avec l'intersection entre eux, il est intéressant d'observer que même si cela réduit la taille de la sélection (dans ce cas, la sélection passe de 11 à 8), les résultats restent les mêmes.

Ainsi, l'union des liens sélectionnés identifiés par la méthode LASSO pour les deux variables qui caractérisent les valeurs quotidiennes du trafic et le modèle de régression linéaire établi avec celles-ci, propose un modèle qui peut estimer les valeurs quotidiennes du réseau avec une erreur moyenne faible et en utilisant seulement 11% des liens du réseau. En ce qui concerne les émissions, le meilleur choix est l'intersection des variables entre elles. Avec seulement 8 liens (3,5% du réseau), toutes les variables peuvent être estimées avec une erreur acceptable, environ 2% pour la distance parcourue, 3% pour les émissions de polluants et moins de 7% pour la vitesse moyenne spatiale.

La méthode LASSO a été utilisée comme méthode de sélection de régression linéaire pour effectuer une sélection des liens les plus pertinents sur le réseau pour les variables de trafic et d'émissions. Pour chacun, un modèle a été établi avec un ensemble de liens et les poids de chacun. L'utilisation

de l'information quotidienne totale/moyenne étant donné que l'apport est suffisant pour estimer avec précision les variables. L'analyse conclut que les liens sélectionnés sur la distance parcourue présentent de meilleurs résultats en termes de nombre minimum de liens nécessaires pour estimer avec précision le trafic et les émissions quotidiennes.

# 4

# Comparaison de l'échantillonnage avec d'autres méthodes statistiques

## 4.1 Introduction

Les techniques statistiques sont couramment utilisées dans de nombreux aspects de la modélisation et de la prévision du trafic. Certaines méthodes exigent un effort de calcul considérable en ce qui concerne le grand nombre d'entrées et leurs variations. D'autres techniques peuvent réduire le temps et l'effort de calcul et améliorer la convergence vers un résultat représentatif. C'est le cas de LASSO présenté dans le chapitre précédent. D'autres méthodes peuvent être utilisées pour des données hautement variables et sont basées sur la probabilité d'occurrence de valeurs correspondantes pour réduire leur variance (Ang and Tang, 2007). Dans ce chapitre, l'accent est mis sur les méthodes statistiques simples qui ont des caractéristiques similaires, telles que la régression linéaire.

Ici, les mêmes données et liens que dans le chapitre précédent sont soumis à d'autres méthodes afin de les comparer. Quatre méthodes ont été choisies : (i) sélection d'échantillonnage aléatoire ; (ii) classement des liens du plus polluant au moins ; (iii) une sélection par étapes qui propose son propre modèle et son processus de sélection optimal ; et (iv) une méthode de partitionnement réseau qui regroupe les liens avant de sélectionner les liens. Pour toutes ces méthodes, l'objectif est d'estimer les émissions journalières ainsi que les valeurs de polluants du réseau de 15 minutes.

La méthode (i) est basique. La méthodologie utilisée est la même que dans une loterie dans laquelle les échantillons sont sélectionnés au hasard. La deuxième méthode est basée sur le classement. Les liens sont classés en fonction de la quantité d'émissions polluantes qu'ils émettent. Les plus polluants sont choisis pour estimer les émissions du réseau en fonction de la régression. La méthode (iii) utilise une régression multiple dans laquelle seuls les membres de la population qui améliorent la précision du modèle sont sélectionnés. Comme LASSO, la méthode par étapes (stepwise) sélectionne les liens les plus pertinents. La dernière méthode, représentée par (iv), envisage d'abord de diviser le réseau en clusters, puis d'utiliser la méthode d'échantillonnage aléatoire appliquée à chaque grappe pour estimer définitivement les valeurs du réseau en fonction de la régression linéaire. Les deux premières techniques d'échantillonnage proposées utilisent une régression linéaire simple pour construire des modèles et estimer les valeurs du réseau. En ce qui concerne les méthodes (iii) et (iv), l'objectif est de comparer d'autres techniques intelligentes pour obtenir des échantillons représentatifs du réseau et éventuellement réduire le taux d'erreurs sur les estimations. La comparaison entre ces modèles peut fournir des informations utiles sur les complexités et l'efficacité variable des modèles d'échantillonnage utilisés.

Pour ce chapitre, seuls deux jeux de données ont été sélectionnés par rapport au dernier chapitre. Le premier, appelé statique / statique, vise à estimer les émissions quotidiennes du réseau à partir d'une émission journalière locale au niveau du lien. Le but du second, appelé statique / dynamique, est identique à celui du jeu de données statique/statique mais avec un niveau temporel raffiné. Ainsi, les émissions du réseau de 15 minutes seront estimées en utilisant les émissions locales de 15 minutes au niveau du lien. Toutes les méthodes ont été appliquées aux deux ensembles de données. Leurs résultats ont été comparés à ceux obtenus avec LASSO et entre eux.

Le chapitre se termine par une description d'une étude utilisant l'emplacement des capteurs réels

installés dans un quartier de Paris. En utilisant leurs emplacements et les valeurs simulées des émissions dans ces endroits, une estimation des émissions du réseau est effectuée avec une régression linéaire appliquée à deux plages temporelles : des estimations quotidiennes et 15 minutes. Les résultats sont analysés et comparés. L'objectif est de quantifier les erreurs en utilisant l'emplacement réel des capteurs lors de l'estimation des émissions au niveau du réseau.

Dans ce résumé seront présentées seulement les conclusions générales du chapitre en comparant les résultats des méthodologies utilisées.

## 4.2 Comparaison de toutes les méthodes

Dans les sections précédentes, plusieurs méthodes ont été proposées pour estimer les émissions du réseau sur deux plages temporelles, quotidiennement et 15 minutes. Toutes ces méthodes ont été comparées à la méthode LASSO détaillée dans le chapitre 3. Les résultats de toutes ces méthodes sont comparés dans cette section.

La première méthode présentée est la sélection aléatoire. Deux méthodologies ont été proposées : la méthode de loterie considérant l'ensemble du réseau et une méthode de loterie à l'intérieur des clusters. Dans les deux cas, le nombre de liens sélectionnés était le même que proposé par LASSO pour chaque jeu de données. Le chapitre 3 a conclu qu'un grand nombre de combinaisons de liens existent pour estimer les émissions du réseau sans affecter l'estimation. Les valeurs estimées pour les jeux de données et les polluants ont eu le même niveau d'erreur que LASSO. L'erreur absolue moyenne pour les 30 sélections de tirage aléatoire a été comparée pour observer la variation de la valeur d'erreur moyenne. La figure 4.1 montre l'erreur moyenne en valeur absolue pour les émissions de $CO_2$ pour chaque échantillonnage aléatoire de chaque ensemble de données.

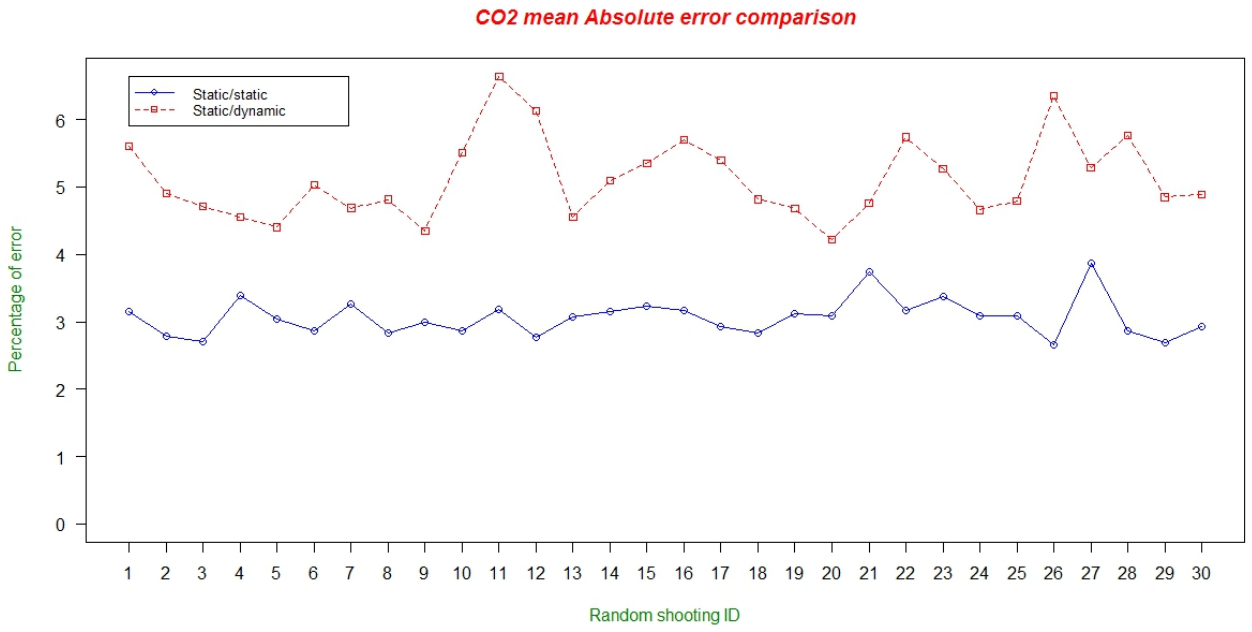



Fig. 4.1 – *L'erreur moyenne en valeur absolue pour chaque échantillonnage aléatoire et jeu de données.*

L'ensemble de données statiques / statiques présente une erreur moyenne plus stable pour toutes les sélections aléatoires que celle statique/dynamique. Pour tous, l'erreur moyenne était d'environ 3%. Compte tenu de l'ensemble de données statiques/dynamiques, les erreurs étaient plus dispersées, entre 4% et 6%, et plus grande que pour le cas statique/statique.

La comparaison des ensembles de données pour les émissions de $NO_x$ présente les mêmes caractéristiques et considérations que pour les émissions de $CO_2$. La seule différence était l'erreur moyenne qui était un peu plus petite dans chaque jeu de données. La figure 4.2 montre l'erreur moyenne en valeur absolue des émissions de $NO_x$ dans les deux jeux de données.

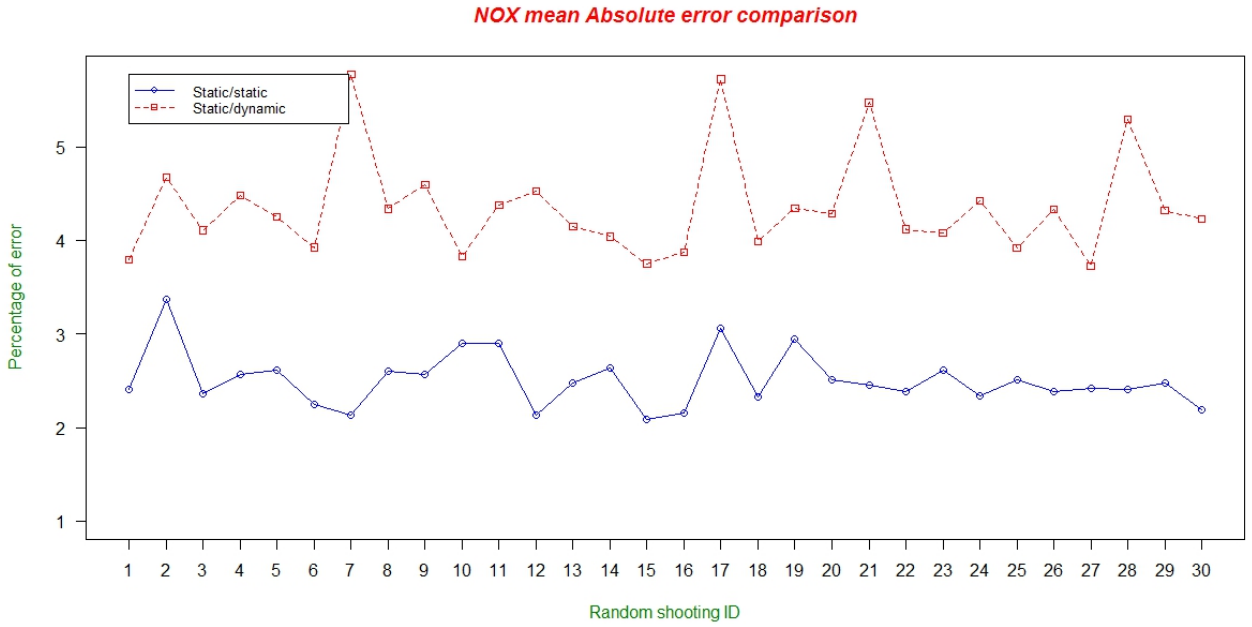



Fig. 4.2 – *L'erreur moyenne en valeur absolue pour chaque échantillonnage aléatoire et jeu de données.*

Réduire l'échelle temporelle des données tend à augmenter le nombre d'erreurs et peut conduire à plus de situations qui ne peuvent être décrites par le modèle (outliers). En outre, le nombre de liens sélectionnés augmente avec la réduction temporelle. Tous les modèles présentés avaient au moins un lien nul sélectionné, ce qui signifie que tous les modèles considéraient moins de liens que ceux sélectionnés par LASSO. LASSO exclut automatiquement tous les liens nuls lors de la construction des modèles, contrairement à la méthode de loterie.

La deuxième méthode était la sélection aléatoire dans les clusters. Cette méthode était basée sur le partitionnement du réseau en petits réseaux qui ont ensuite été analysés itérativement pour définir les clusters en fonction de leurs valeurs d'émission. L'objectif était de réduire le réseau en quelques clusters en fonction des valeurs d'émissions, afin de rassembler tous les liens ayant un comportement d'émission similaire dans le même cluster. Une fois les clusters définis, une sélection aléatoire a été effectuée au même rythme que LASSO en fonction de la taille du échantillon. Cette approche réduit les erreurs d'estimation des émissions sur les liens sélectionnés qui appartiennent à un cluster et qui peuvent être plus représentatifs de l'ensemble du groupe en raison de leur similitude. Cette méthode permet de sélectionner des liens pour représenter les clusters et estimer les émissions du réseau.

En général, les deux polluants ont les mêmes partitions de cluster dans chaque ensemble de données, ce qui est positif car cela signifie que les deux polluants peuvent être estimés avec la même partition. En autre, dans tous les cas, le regroupement réduit considérablement la répartition des erreurs par rapport à LASSO. Aussi, tous les modèles sont plus petits que LASSO en raison des liens nuls qui peuvent être sélectionnés par la méthode aléatoire. Cette tendance est confirmée si tous les clusters sont représentés dans le modèle final. En ce qui concerne le cas de $NO_x$ dans le jeu de données statiques / statiques, le modèle donne des erreurs plus importantes pour le réseau 6-cluster que LASSO car les liens sélectionnés dans les clusters étaient de liens à valeur nulle.

La troisième méthode présentée était les liens classés. L'idée principale est de classer les valeurs d'émission des liens du réseau du plus au moins polluant. Les 11 liens sélectionnés représentent en moyenne 21% des émissions du réseau pour le trafic quotidien. Les erreurs du modèle dans les liens classés sont moins dispersées que LASSO. Cet ensemble de données a presque 7 fois moins de prédicteurs dans le modèle que l'ensemble de données statiques/dynamiques. Lorsque la plage temporelle est réduite, la variabilité sur les données est incorporée et, par conséquent, le modèle nécessite plus de prédicteurs pour obtenir une bonne estimation des émissions. La taille du modèle de dataset statique/dynamique est égale à 77 prédicteurs pour $CO_2$ et 65 pour $NO_x$. Le modèle basé sur les liens classés a moins d'erreurs dispersées que le modèle optimal de LASSO. Les liens classés pour $CO_2$

représentent 76% des émissions totales du réseau et les liens classés de $NO_x$ représentent, en moyenne, 70% des émissions du réseau pour une période de 15 minutes.

La quatrième et dernière méthode est la méthode par étapes (stepwise). C'est une méthode intelligente comme LASSO. Pour tous les cas, la méthode par étapes a sélectionné moins de liens et les erreurs étaient moins dispersées que pour la méthode LASSO.

Pour mesurer l'efficacité des modèles, et choisir le meilleur parmi eux, le critère Bayesian d'information $BIC$ (Schwarz, 1978) a été utilisé pour sélectionner le modèle qui correspond le mieux aux données. Le score $BIC$ a été calculé pour chaque modèle construit pour les émissions $CO_2$ et $NO_x$ dans les deux jeux de données. Table 4.1 montre le score $BIC$ pour tous les modèles étudiés.

| | Static/static | | Static/dynamic | |
|---|---|---|---|---|
| VARIABLES → / MODELS ↓ | $CO_2$ | $NO_X$ | $CO_2$ | $NO_X$ |
| Reference | 5469 | 3880 | 128111 | 90488 |
| LASSO 1SE | 5341 | 3885 | 115976 | 78197 |
| Random ID 2 | - | 3817 | - | - |
| Random ID 3 | 5338 | - | - | - |
| Random ID 7 | - | - | - | 78917 |
| Random ID 14 | 5363 | - | - | - |
| Random ID 15 | - | 3742 | - | - |
| Random ID 16 | - | - | 116580 | - |
| Random ID 19 | - | - | - | 78499 |
| Random ID 20 | - | 3766 | 116010 | - |
| Random ID 26 | - | - | 116779 | - |
| Random ID 27 | 5401 | - | - | 78241 |
| 4 clusters | 5382 | 3751 | 117414 | 79275 |
| 6 clusters | 5343 | 3854 | 117375 | 79048 |
| Ranked links | 5345 | 3740 | 115989 | 78118 |
| Stepwise | 5328 | 3746 | 115871 | 78034 |

Tab. 4.1 – *Le scores BIC calculé pour chaque modèle d'émission.*

Les scores $BIC$ sont similaires pour les modèles $CO_2$ utilisant l'ensemble de données statiques / statiques. La méthodologie $BIC$ indique que le modèle avec le score le plus bas est le meilleur pour les données, et dans ce cas, ce modèle est la sélection par étapes. L'ID Random 3 a le deuxième score $BIC$ le plus bas, suivi du modèle LASSO 1SE et des modèles "6-cluster" et des classements classés. Compte tenu des émissions de $NO_x$ dans le même ensemble de données, le modèle de lien classé a le score le plus bas suivi des modèles Random ID 15 (le meilleur modèle de tous les tirages aléatoires) et ensuite par le modèle le stepwise.

Compte tenu de l'ensemble de données statiques / dynamiques, pour les deux émissions polluantes, le modèle par étapes a le score $BIC$ le plus bas, suivi du LASSO 1SE et des modèles de liens classés. Il est intéressant de noter que, pour tous les modèles proposés à la fois pour les polluants et les deux

jeux de données, la méthode par étapes a presque toujours le meilleur score $BIC$ ou est parmi les trois meilleurs scores par rapport aux autres modèles. Les deux autres modèles avec des scores similaires pour cet ensemble de données sont le LASSO 1 SE et le modèle de liens classés. Comme on peut le constater, le meilleur score $BIC$ est obtenu par deux méthodes d'échantillonnage intelligentes et par une méthode naïve. Les distributions d'erreurs estimées sont assez similaires entre ces trois modèles, même avec différentes tailles de modèles. La méthode naïve considère uniquement les liens les plus polluants du réseau avec le même taux de sélection que LASSO, 77 liens pour $CO_2$ et 65 liens pour $NO_x$. Cette méthode peut estimer les valeurs du réseau dans n'importe quelle plage dynamique du réseau pour un faible coût de calcul et elle est facile à utiliser. Mais le modèle par étapes nécessite moins de liens dans le réseau pour le même coût de calcul et il obtient une estimation avec une erreur associée moins dispersée que les deux autres.

En général, pour les deux polluants, les modèles de tirage aléatoire et les meilleurs modèles de tirage aléatoire appliqués à chaque polluant ont des scores similaires de $BIC$. En outre, le lien classé et les modèles LASSO 1SE sont parmi les meilleurs scores possibles. De plus, comme le montre ce chapitre, les conclusions montrent que la méthode de clustering est la même que celle des meilleures estimations d'émission pour différentes échelles, malgré son coût de calcul.

# 5

# Effet de l'agrégation des données de trafic sur les estimations d'émission

## 5.1 Introduction

L'objectif de cette section est d'examiner l'influence de l'agrégation spatiale et temporelle sur les variables de trafic utilisées pour estimer les émissions polluantes des fonctions COPERT. Les données simulées du 6$^{\text{ème}}$ arrondissement de Paris ont été utilisées pour explorer ces relations. Initialement, deux échelles d'agrégation spatiale et temporelle ont été proposées. En ce qui concerne l'agrégation spatiale, l'échelle locale est définie par les variables de trafic au niveau du lien tandis que l'échelle globale tient compte de l'ensemble du quartier. Dans les deux cas, l'objectif est d'estimer les émissions au niveau du réseau. Compte tenu de l'agrégation temporelle, les données de trafic peuvent être récupérées directement par les capteurs toutes les 15 minutes et être également regroupées pour qualifier le trafic quotidien. Ces deux échelles appliquées au cas de Paris permettent d'effectuer une analyse de sensibilité pour déterminer dans quelle mesure la gamme possible de ces variables d'entrée influence les résultats du modèle (c.-à-d. les émissions de $CO_2$ et $NO_x$ pour les tronçons routières) et donc la précision. La précision des estimations d'émission peut être affectée par les erreurs associées au modèle d'émission lui-même (c'est-à-dire les facteurs d'émission) et les erreurs associées aux variables d'entrée dans le modèle d'émission. L'évaluation des émissions dues au trafic utilise couramment des données d'entrée mesurées pour des variables clés telles que les kilomètres de véhicules parcourus (c.-à-d. le volume de trafic multiplié par la longueur de la route), la vitesse moyenne et la composition du mélange de flotte pour valider la méthode présentée dans (Pierson et al., 1996) et (Mukherjee and Viswanathan, 2001) . Par conséquent, ces études ont tendance à quantifier les erreurs associées uniquement au modèle d'émission, afin de valider la méthodologie appliquée mais n'évaluent pas directement l'exactitude des estimations couplées des émissions de trafic. De même, l'extension de la zone de validation à une zone supérieure à une route permet d'étudier et de quantifier les erreurs associées, car les estimations sont basées sur la combinaison de modèles de trafic et d'émissions. L'objectif est de quantifier les erreurs et de souligner qu'un biais peut être identifié. Cela montre que de tels calculs manquent de cohérence entre échelles, en particulier lorsque la dynamique du trafic et les congestions sont prises en compte correctement. Par la suite, les fonctions COPERT sont analysées pour comprendre comment elles ont été construites et comment elles peuvent être utilisées correctement pour estimer les émissions. L'objectif est de montrer et d'identifier la source des biais d'estimation et leur quantification. Nous montrons également quelle échelle est la plus précise pour estimer les émissions.

## 5.2 Le quantification des émissions

Les données simulées décrites dans le chapitre 2 ont été utilisées comme base de cette étude. Comme dans les chapitres précédents, les fonctions d'émission de COPERT et les données de trafic pour estimer les émissions au niveau du réseau ont été utilisées. À cette fin, deux échelles de couplage principales ont été définies : (i) les émissions sont calculées pour tous les liens en fonction des distances

de déplacement et de la vitesse moyenne pour chaque période de temps, puis agrégées pour déterminer les émissions quotidiennes du réseau et (ii) les variables de trafic sont d'abord agrégées au niveau du réseau, puis les émissions sont dérivées en fonction des valeurs agrégées. Pour expliquer, la première méthode appelée "échelle locale", les variables de trafic sont récupérées toutes les 15 minutes dans chaque lien, puis les émissions sont calculées directement avec ces valeurs. Ensuite, toutes les valeurs d'émission au niveau du lien sont recueillies pour estimer les valeurs totales des émissions quotidiennes au niveau du réseau. La deuxième méthode de calcul, appelée échelle globale, commence à rassembler toutes les valeurs des variables de trafic au niveau du lien du réseau, puis les émissions sont calculées en fonction de ces valeurs. Ces deux échelles spatio-temporelles sont identifiées pour comparer les résultats en terme d'émission obtenues à partir des mêmes données de trafic mais agrégées différemment. Pour chaque échelle, l'impact sur la sortie du modèle est ensuite évalué pour tester l'influence de l'agrégation sur de multiples simulations avec différents paramètres.

Les émissions utilisant la méthodologie COPERT sont déterminées par trois variables principales, à savoir la distance parcourue (c'est-à-dire le volume de trafic multiplié par la longueur du lien), la vitesse moyenne et la composition du trafic. La quantité de voyage a un effet important sur les estimations d'émission, en particulier parce que toute erreur dans les distances parcourues totales par les voitures est propagée proportionnellement dans les estimations d'émission. Par rapport aux autres variables telles que la vitesse moyenne et la composition de la flotte, les distances parcourues exactement sur les rues sont faciles à obtenir en tant que données de compte de trafic qui sont mesurées en différents points (par exemple détection de capteurs, caméras, relevé de comptage manuel) dans les réseaux routiers. À partir du modèle microscopique dynamique, toutes les valeurs exactes des distances parcourues par les voitures ont été récupérées à partir de chaque liaison, et étant donné qu'il s'agit d'une variable linéaire, elle ne présente peu de différence entre les deux échelles proposées.

Les entrées de vitesse imprécises peuvent avoir un effet important sur les émissions estimées (Smit, 2008). (Chatterjee et al., 1997) ont effectué une analyse de sensibilité sur le modèle d'émission qu'utilise la vitesse moyenne et montrent qu'une erreur de 5 km/h dans la valeur de la vitesse moyenne utilisée comme entrée dans le modèle d'émission pour une autoroute a provoqué une différence de 42% dans l'estimation des émissions de $CO_2$ en raison d'une relation fortement non linéaire entre les courbes des facteurs d'émission et la vitesse moyenne. Les deux échelles définies, locales et globales, sont appliquées ici. La différence de valeurs de vitesse moyenne par définition a été explorée dans chaque échelle, ainsi que sur la façon dont les répercussions sur les estimations d'émission.

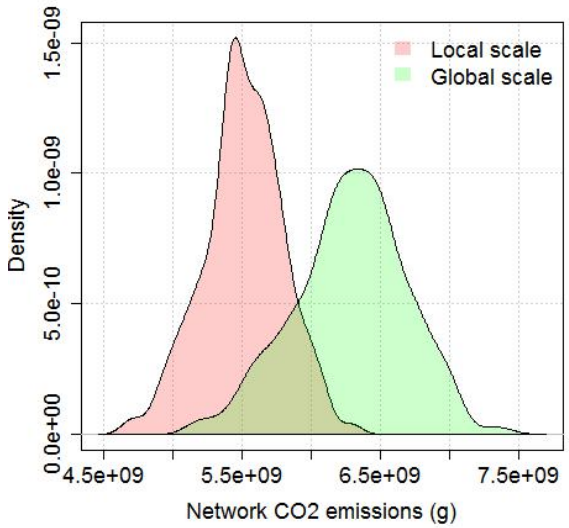
Compte tenu des données de trafic microscopiques, des facteurs d'émission pour les émissions à chaud d'échappement et de la composition du parc automobile français, les émissions de $CO_2$ et $NO_x$ ont été calculées à la fois dans les échelles proposées, locales et globales et elles sont indiquées dans la figure 5.1.

La figure 5.1 montre les émissions quotidiennes de polluants du réseau (c.-à-d. 400 mesures avec différentes conditions de circulation). Comme on peut le constater, les résultats de la quantification des émissions dans les deux échelles sont complètement différents. Pour les deux polluants, les émissions calculées avec des données à l'échelle globale ont des valeurs moyennes plus élevées, et les données sont plus dispersées que les échelles locales.

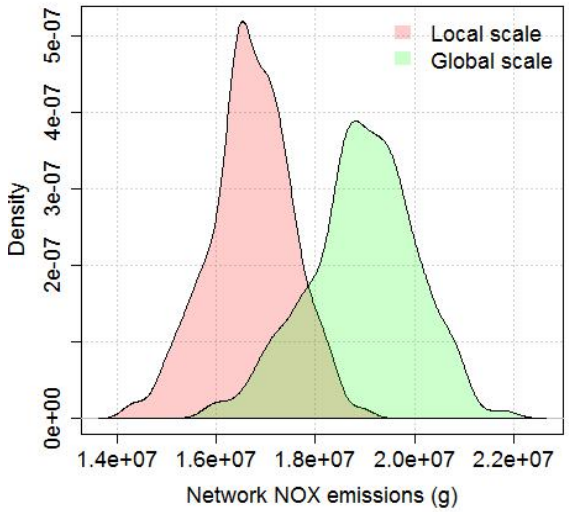
Ces résultats montrent que l'utilisation des fonctions d'émission COPERT à différentes échelles spatio-temporelles induit un biais en fonction de l'agrégation des données de trafic principalement lorsque l'on considère la dynamique du trafic, c'est-à-dire la variation de temps de la vitesse. Les mêmes données, qui représentent plus d'un an de trafic journalier et différents niveaux de demande, agrégées à l'échelle locale (lien) et à l'échelle globale (réseau) fournissent une quantification différente des polluants.

La courbe de facteurs d'émission n'est pas une fonction linéaire, ce qui signifie que les valeurs entre les échelles seront différents en raison de la convexité des fonctions d'émission. Le degré de convexité dépend de chaque polluant.

Après avoir analysé en profondeur d'où vient la source de ce biais, on a conclu que le biais entre les échelles globale et locale peut être estimé à partir de trois termes : le premier terme définit l'écart entre les fonctions d'émission en fonction des échelles locales et globales. Le deuxième terme correspond à la convexité de la fonction en considérant f> 0, et le dernier terme correspond aux corrélations entre les




(a) La distribuition de la densité des émissions de $CO_2$ dans le réseau.




(b) La distribuition de la densité des émissions de $NO_x$ dans le réseau.

Fig. 5.1 – *La distribution de la densité de émissions de polluants dans le réseau en utilisant les approches locales et globales.*

variables, c'est-à-dire les distances totales parcourues et les émissions unitaires définies par la vitesse moyenne. Cela met en évidence le fait que le biais entre les échelles dépend fortement de la vitesse moyenne. En raison de sa nature non linéaire, la vitesse moyenne ne sera pas la même dans toute agrégation spatiale et temporelle de la même population.

Dans notre étude de cas, l'approche globale tend à surestimer les émissions d'environ 14% en moyenne par rapport à l'échelle locale. En ce qui concerne les émissions de polluants, lorsque l'on considère le dioxyde de carbone (c'est-à-dire les émissions de $CO_2$), la surestimation peut être comprise entre 10% et 18%, alors que pour les oxydes d'azote (c'est-à-dire les émissions de $NO_x$), la surestimation est comprise entre 11% et 16% . En considérant les conditions de circulation, les disparités sont d'environ 4% pour $CO_2$ et 7% pour les émissions de $NO_x$ en situation fluide. Dans une situation congestionnée, la surestimation peut atteindre environ 23% par rapport à l'approche locale.

## 5.3 Conclusions

La qualité des données sur les émissions de trafic est évidemment un facteur important pour la précision des estimations d'émission. Comme l'a souligné (Smit, 2008), les données d'entrée les plus pertinentes pour l'évaluation des émissions devraient être identifiées. Le même auteur a également abordé l'importance d'assurer l'exactitude et la validation du modèle d'émission de trafic couplé pour améliorer encore le développement des inventaires des émissions.

L'étude couverte par ce travail pourrait être utile pour évaluer l'ampleur des données de trafic réel utilisées par les gestionnaires de villes lors de l'estimation des contributions des émissions de trafic routier et de l'évaluation des stratégies pour les abaisser.

# 6

# Conclusion générales et perspectives

L'objectif de cette thèse était de définir des méthodes d'échantillonnage pour estimer les émissions totales à plusieurs échelles spatiales et temporelles à partir d'un échantillon. La motivation de cette étude est d'améliorer la précision des données d'entrée nécessaires dans les modèles d'émissions afin de pouvoir quantifier les émissions routières de manière fiable. Cette précision est liée à la dynamique du trafic, car les volumes d'émissions polluantes sont sensibles à plusieurs facteurs tels que le comportement du conducteur (agressivité), les conditions de trafic (free-flow ou congestion), le mode de fonctionnement du véhicule (accélérations, décélérations et stop-and-go), le parc automobile, le type d'infrastructure routière, etc. En réalité, ces informations peuvent provenir d'enquêtes et d'observations d'études sur le terrain ou de technologies routières telles que les capteurs électromagnétiques ou caméras. Ces méthodologies sont extrêmement coûteuses et ne peuvent représenter le réseau qu'au niveau local. Les simulations de trafic microscopiques sont de plus en plus utilisées pour mieux évaluer les réseaux de transport. Elles permettent d'évaluer, de tirer des conclusions et de tester de nouvelles techniques sans avoir besoin de perturber les systèmes réels et d'effectuer de nouvelles collectes des données. Elles fournissent également des informations sur le trafic décrites dans le temps et dans l'espace, ce qui constitue le grand avantage des modèles de simulation.

Cependant, la simulation microscopique du trafic d'un réseau peut prendre beaucoup de temps pour traiter le volume d'informations, ce qui n'est pas toujours aisé sur le plan informatique : c'est précisément le point sur lequel se concentre cette thèse. L'échantillonnage a été proposé pour réduire le volume de données à traiter tout en maintenant la précision et la fiabilité. En utilisant les méthodes statistiques existantes, au lieu d'estimer les émissions pour chaque lien dans le réseau, l'idée est d'identifier les liens les plus représentatifs du réseau pour estimer les émissions globales. L'identification des bonnes méthodes statistiques qui tiennent compte des corrélations spatio-temporelles des données de trafic est importante pour la représentativité de l'échantillon. L'état du trafic et sa dynamique évoluent dans le temps et l'espace et sont liés aux changements de demande dans le réseau ainsi qu'aux périodes de pointe. À cette fin, il est nécessaire d'obtenir une bonne connaissance du couplage entre les modèles de trafic et d'émission.

En ce qui concerne les modèles d'émissions, les modèles statiques sont adaptés aux études stratégiques, comme la préparation des inventaires des émissions dans les principaux domaines où les conditions moyennes sont satisfaisantes. La méthodologie COPERT a été identifiée comme facilitant l'applicabilité du couplage pour les objectifs proposés dans la thèse et aussi pour considérer leur application dans des cas réels. Ce modèle est basé sur la mesure des émissions pendant les cycles représentatifs et couvre tous les polluants atmosphériques des différentes catégories de véhicules et sont représentés par les facteurs d'émission.

Grâce aux diverses études réalisées, ce travail m'a permis d'identifier les enjeux associés à l'estimation des émissions au niveau urbain. Parmi les études proposées, l'importance et l'influence des approches de trafic agrégées et désagrégées, en ce qui concerne le couplage du trafic et des modèles d'émissions, se sont révélées être un point essentiel pour assurer la fiabilité des résultats. Les principales précautions à prendre en compte lors du couplage sont la cohérence des représentations du réseau routier, les flux et les conditions de circulation dynamiques. L'étalonnage des modèles était

également d'une grande importance pour assurer la qualité des résultats. En outre, il est important de souligner que les modèles d'émissions statiques tels que COPERT, qui estime les émissions basées sur les facteurs d'émission, peuvent également présenter une fausse représentation des émissions lorsqu'on considère l'ampleur de l'application. Ceci est dû à la différence entre les cycles de conduite normalisés intégrés dans le modèle et les cycles de conduite réels, ce qui entraîne des niveaux d'émission biaisés.

Il est important de souligner que l'objectif de ce travail n'était pas d'évaluer les valeurs absolues des émissions mais d'identifier l'influence et la variabilité des estimations en raison des différences dans l'application des modèles et des données d'entrée. Cependant, il est important de noter que les facteurs d'émission utilisés par les modèles d'émission dans cette thèse devraient être basés sur les mesures du véhicule dans la zone d'étude.

En ce qui concerne la dimension opérationnelle, le défi consiste donc à améliorer la qualité de l'air sans compromettre la mobilité de la population. Confrontée au problème des données manquantes et à la difficulté d'obtenir une représentation complète du trafic d'un réseau de transport complet, la méthodologie d'échantillonnage peut aider à choisir la stratégie de trafic la plus efficace à adopter pour chaque type de problème traité. Elle peut également évaluer la performance des stratégies de gestion du trafic pour réduire les émissions des polluantes. Les applications de ces techniques sont nombreuses. En plus d'une amélioration significative du temps de calcul, l'élaboration de méthodes d'échantillonnage appropriées pourrait également aider à identifier les domaines clés d'un réseau, à améliorer les évaluations a posteriori (positionnement optimal des instruments de mesure tels que les capteurs sur route, la définition des voyages des véhicules de référence avec mesures à bord). Les méthodologies couvertes par ce travail pourraient également être utiles dans les évaluations de la quantification des émissions en temps réel.

La pertinence du couplage des modèles de trafic et d'émission alliés à la méthode d'échantillonnage peut également être renforcée par le fait qu'ils sont une aide importante dans le processus de prise de décision, car ils peuvent fournir de grandes quantités de données de manière systématique et reproductible, ce qui permet une comparaison a priori de plusieurs scénarios alternatifs sans qu'il soit nécessaire de les mettre en oeuvre. De cette façon, la méthodologie offre des avantages opérationnels importants lorsqu'elle est appliquée en tant qu'outil d'évaluation, mais il faut tenir compte de ses inconvénients.

La méthode développée dans cette thèse présente un grand potentiel pour la formulation de diagnostics d'émission dans les zones urbaines et pour l'évaluation de différents scénarios de performance routière. Enfin, la méthode proposée dans ce travail peut stimuler d'autres études sur ce sujet, car des études spécifiques conduiront à des résultats qui peuvent contribuer à son amélioration.

En tant que recommandation pour les travaux futurs, l'application de la méthode proposée est proposée dans différents contextes, impliquant des zones urbaines avec différentes infrastructures routières et des caractéristiques opérationnelles et des tailles de réseaux et des hétérogénéités. En outre, bien que cette recherche se limite à étudier uniquement les voitures particulières et les émissions d'échappement chaudes, différentes compositions de flottes de véhicules et types de carburants peuvent être incorporés dans la méthode. En outre, le couplage du modèle de trafic microscopique avec d'autres modèles d'émissions serait rentable et serait un moyen de vérifier l'efficacité et l'efficience de la méthodologie pour différents réseaux et de vérifier si cela peut être utilisé de cette façon.

En effet, les réseaux de capteurs utilisés pour mesurer la pollution atmosphérique sont généralement de faible densité et ne discriminent pas spécifiquement la contribution des émissions de trafic routier. À l'inverse, les données de trafic en temps réel et surtout les véhicules à sondage se développent énormément. Une idée intéressante serait donc de compléter les mesures par des capteurs dédiés avec des informations provenant de véhicules traceurs couplés à un modèle d'émission. Encore une fois, l'identification des informations de trafic pertinentes par des méthodes d'échantillonnage appropriées prendrait tout son sens.

# Bibliographie


Ang, A. and Tang, W. (2007). *Probability concepts in engineering*. John Wiley and Sons Inc.

Burghout, W. (2004). Hybrid microscopic-mesoscopic traffic simulation. Doctoral thesis, Royal Institute of Technology, Stockholm.

Can, A. and Leclercq, L. (2009). Estimation des consommations énergétiques et des polluants émis par le trafic routier : Revue bibliographique des modèles existants. Technical report, IFSTTAR.

Cappiello, A. (2002). Modelling traffic flow emissions. Master of science in transportation, Massachussetts institute of technology.

Carslaw, D. C. and Beevers, S. D. (2005). Estimation of road vehicle primary no2 exhaust emission fractions using monitoring data in london. *Atmospheric Environmental*, 39 :167–177.

Chatterjee, A., Miller, T., Philpot, J., Wholley, T., Guensler, R., Hartgen, D., Margiotta, R., and Stopher, P. (1997). Improving transportation data from mobile source emission estimates. In *NCHRP report No. 394 on Transportation Research Boarder*.

De Nevers, N. (2000). *Air pollution control engineering*. New York : McGraw-Hill, 2 edition.

DOE (2015). Australia's emissions projections. Report 2014-15, departament of the Environment.

EEA (2016). Use of cleaner and alternative fuels. Online URL : https ://www.eea.europa.eu/data-and-maps/indicators/use-of-cleaner-and-alternative-fuels/use-of-cleaner-and-alternative-10, Accessed July 10, 2017.

EEA (2017a). Size of the vehicle fleet in eu. Online URL : https ://www.eea.europa.eu/data-and-maps/indicators/size-of-the-vehicle-fleet/size-of-the-vehicle-fleet-7, Accessed July 09, 2017.

EEA (2017b). Transport emissions of greenhouse gases in eu. Online URL : https ://www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-10, Accessed July 09, 2017.

EU (2013). Environmental action programme to 2020. Online URL : http ://ec.europa.eu/environment/action-programme/, Accessed July 09, 2017.

Font, A. and Fuler, G. W. (2016). Did policies to abate atmospheric emissions from traffic have a positive effect in london ? *Environmental Pollution*, 218 :463–474.

Frecht, D., Fischer, P., Fortunato, L., Hoek, G., De Hoogh, K., Marra, M., Kruize, H., Vienneau, D., Beelen, R., and Hansell, A. (2015). Associations between air pollution and socioeconomic characteristics, ethnicy and age profile of neighborhoods in england and the netherlands. *Environmental Pollution*, 198 :201–210.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1) :1.

GLA (2010). Clearing the air : the mayor's air quality strategy. Report 176, Greater London Authority.

IPCC (2013). Climate change 2013 : The physical science basis. The fifth assessment report of the intergovernmental panel on climate change, Cambridge University.

Jie, L., Zuylen, H. V., Chen, Y., Viti, F., and Wilmink, I. (2013). Calibration of a microscopic simulation model for emission calculation. *Transportation Research Part C*, 31 :172–184.

Mukherjee, P. and Viswanathan, S. (2001). Carbon monoxide modeling from transportation sources. *Chemosphere*, 45 :1071–1083.

Pierson, W., Gertler, A., Robinson, N., Sagebiel, J., Zielinska, B., Bishop, G., Stedman, D., Zweidinger, R., and Ray, W. (1996). Real-world automotive emissions - summary of studies in the fort mchenry and tuscarora mountain tunnels. *Atmospheric Environment*, 30 :2233–2256.

Schiper, N., Lejri, D., and Leclercq, L. (2016). How selection techniques on traffic data sets can help in estimating network vehicle emissions. *Journal of Earth Sciences and Geotechnical Engineering*, 6 :51–69.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2).

Smit, R. (2008). Errors in model predictions of nox traffic emissions at road level - impacts of input data quality. *WIT Transactions on Ecology and the Environment*, 116 :255–269.

Sturm, P. J., Pucher, k., and Sudy, C. Almbauer, R. A. (1996). Determination of traffic emissions – intercomparison of different calculation methods. *The Science of Total Environment*, 189 :187–196.

TfL (2014). Travel in london. Report 7, Transport for London.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58 :267–288.

Tsanakas, N., Ekstrom, J., and Olstam, J. (2017). Reduction of errors when estimating emissions based on static traffic model outputs. *Transportation Research Procedia*, 22 :440–449.

UNFCCC (2012). Kyoto protocol. Online URL : http ://unfccc.int/kyotoprotocol/items/2830.php, Accessed July 10, 2017.

UNFCCC (2014). United nations framework convention on climate changes. Online URL : http ://unfccc.int/2860.php, Accessed July 09, 2017.

Villegas, D., Canaud, M., and Bécarie, C. (2013). Constitution d'un environnement de simulation à grande échelle et production de données synthétiques pour le trafic routier et étude des données mobiles pour l'analyse des déplacements piétons- ISpace&Time. Technical report D4.4, IFSTTAR, Paris.

WHO (2013). Review of evidence on health aspects of air pollution. Technical Report REVIHAAP first results, World Health Organization.

FOLIO ADMINISTRATIF

THÈSE SOUTENUE DEVANT L'ÉCOLE NATIONALE DES TRAVAUX PUBLICS DE L'ÉTAT

**NOM :** SCHIPER **DATE DE SOUTENANCE PREVUE :** 09/10/2017
(avec précision du nom de jeune fille le cas échéant)

**Prénoms :** Nicole,

**TITRE :** Échantillonnage des données de trafic pour l'estimation de la pollution atmosphérique aux différentes échelles urbaines

**École doctorale :** Mécanique, Énergétique, Génie civil, Acoustique (MEGA)

**NATURE :** Doctorat **Numéro d'ordre :**

**Spécialité :** Génie civil

**Code B.I.U.** - / et bis **CLASSE :**

**RÉSUMÉ :** La circulation routière est une source majeure de pollution atmosphérique dans les zones urbaines. Les décideurs insistent pour qu'on leur propose de nouvelles solutions, y compris de nouvelles stratégies de management qui pourraient directement faire baisser les émissions de polluants. Pour évaluer les performances de ces stratégies, le calcul des émissions de pollution devrait tenir compte de la dynamique spatiale et temporelle du trafic. L'utilisation de capteurs traditionnels sur route (par exemple, capteurs inductifs ou boucles de comptage) pour collecter des données en temps réel est nécessaire mais pas suffisante en raison de leur coût de mise en œuvre très élevé. Le fait que de telles technologies, pour des raisons pratiques, ne fournissent que des informations locales est un inconvénient. Certaines méthodes devraient ensuite être appliquées pour étendre cette information locale à une grande échelle. Ces méthodes souffrent actuellement des limites suivantes : (i) la relation entre les données manquantes et la précision de l'estimation ne peut être facilement déterminée et (ii) les calculs à grande échelle sont énormément coûteux, principalement lorsque les phénomènes de congestion sont considérés. Compte tenu d'une simulation microscopique du trafic couplée à un modèle d'émission, une approche innovante de ce problème est mise en œuvre. Elle consiste à appliquer des techniques de sélection statistique qui permettent d'identifier les emplacements les plus pertinents pour estimer les émissions des véhicules du réseau à différentes échelles spatiales et temporelles. Ce travail explore l'utilisation de méthodes statistiques intelligentes et naïves, comme outil pour sélectionner l'information la plus pertinente sur le trafic et les émissions sur un réseau afin de déterminer les valeurs totales à plusieurs échelles. Ce travail met également en évidence quelques précautions à prendre en compte quand on calcul les émissions à large échelle à partir des données trafic et d'un modèle d'émission. L'utilisation des facteurs d'émission COPERT IV à différentes échelles spatio-temporelles induit un biais en fonction des conditions de circulation par rapport à l'échelle d'origine (cycles de conduite). Ce biais observé sur nos simulations a été quantifié en fonction des indicateurs de trafic (vitesse moyenne). Il a également été démontré qu'il avait une double origine : la convexité des fonctions d'émission et la covariance des variables de trafic.

**MOTS-CLES :** *Émissions de véhicules*, *Échantillonage de données de trafic*, *Agrégation spatio-temporelle*, *Échelles de réseau*, *Modèle d'émission*.

**Laboratoire(s) de recherche :** Laboratoire d'Ingénierie Circulation Transport (LICIT)

**Directeur de thèse :** Ludovic Leclercq

**Président de jury :** Christine SOLNON (Université de Lyon)

**Composition du jury :** Latifa OUKHELLOU (Université de Marne-la-Vallée), Rapporteur
Nikolaos GEROLIMINIS (École polytechnique fédérale de Lausanne), Rapporteur
Delphine LEJRI (Université de Lyon), Encadrante
Ludovic LECLERCQ (Université de Lyon), Directeur de thèse