atmosphere

Supplementary materials: Routes Alternatives with Reduced Emissions: Large-Scale Statistical Analysis of Probe Vehicle Data in Lyon

Alexandre Jayol *, Delphine Lejri and Ludovic Leclercq

Univ Lyon, Univ Eiffel, ENTPE, LICIT-ECO7, F-69675 Lyon, France
* Correspondence: alexandre.jayol@entpe.fr; Tel.: +33-472-047-165

Table S1. Table of abbreviations

Abbreviation	Interpretation
CAN	Controller Area Network
$card(v_{obs})$	Total number of mean speed flows observed for November at each period
$card(\Omega)$	Total number of mean speed flows expected
CMEM	Comprehensive Modal Emission Model
DiGraphs	Directed Graphs
ΔTE	Delta Travel Emission
ΔTD	Delta Travel Distance
ΔTT	Delta Travel Times
d^l	distance of the link
DR	Detour ratio
DR_{ED}	Detour Ratio method based on Euclidean Distance
DR_{SP}	Detour Ratio method based on the shortest distance path
DR_{SuP}	Detour Ratio method induced by Sustainable Path
ECDF	Empirical Cumulative Distribution Function
ECO-AFA	ECO-Agent Feedback Assignment
ED	Euclidean Distance of the OD
EF	Emission Factor
f^k	COPERT emission factor (g/km) of pollutant k
f^k_{scaled}	COPERT scaled emission factor (g/km) of pollutant k
$f(v_{obs})$	frequency of the mean speed flow observed
FC	Fuel Consumption
FCD	Floating Car Data
ffs	free-flow speed
fow	form of way
FP	Fastest Path
frc	functional road class
HBEFA	Handbook Emission Factors for Road Transport
LASSO	Least Absolute Shrinkage and Selection Operator
netclass	Relevance in the network
OD	Origin-Destination
OEMs	On-road Emission Measurements
PEMS	Portable Emission Measurement System
PHEM	Passenger car and Heavy duty Emission Model
R^2	Correlation coefficient
RMSE	Root Mean Square Error
RF	Random Forest
SP	Shortest Path
std	Standard deviation
SUV	Sport Utility Vehicle
SVM	Support Vector Machine
TD	Travel Distance
TD_{actual}	TD of the actual trip
$TD_{alternative}$	TD of the alternative trip
TD_{SP}	TD of the shortest Path between the OD of the actual trip considered
TE	Travel Emission
TS	Travel Speed
TT	Travel Time
TT_{actual}	Travel Time of the actual trip

$TT_{alternative}$	Travel Time of the alternative trip
TT-AFA	Travel Time-Agent Feedback Assignment
V_{link}	Mean speed flow observed or deduced by RF of a link
v_{obs}	Mean speed flow observed of a link
VSP	Vehicle Specific Power

Table S2. Links frc distribution.

Category	Connotation	Count
0	Motorway, Freeway, or Other Major Road	325
1	a Major Road Smaller than a Motorway	976
2	Other Major Road	1 931
3	Secondary Road	4 533
4	Local Connecting Road	1 909
5	Local Road of Great Importance	5 089
6	Local Road	8 033
7	Local Road of Minor Importance	17 016

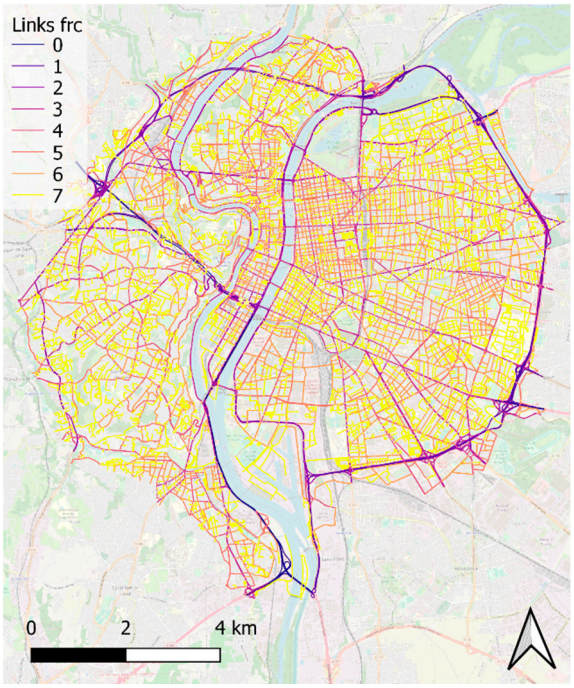


Figure S1. Spatial distribution of frc links.
(*background map: © OpenStreetMap*)

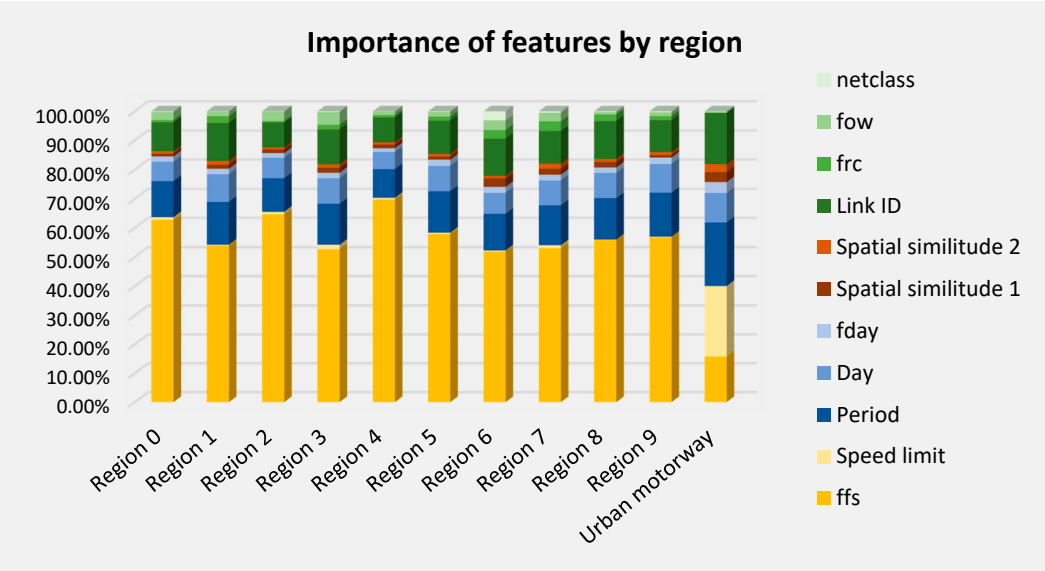


Figure S2. Representation of the importance of the features according to the RF model of each region.

Point S1

Mean speed flow

In traffic theory, the right way to compute the mean speed is to consider TT and TD because these variables are additive and transferable at different scales. For a given network link l of length d^l and with n vehicles observed for a period p, we denote $d_i^{l,p}$ the distance traveled by a vehicle i ($i = [\![1:n]\!]$) during an interval $\Delta t_i^{l,p}$ with a coverage $c_i^{l,p}$. Thus, the mean speed of the link l considered $V^{l,p}$ is computed for a period p as follows in Equation (S1):

$$V^{l,p} = \frac{\sum_i^n d_i^{l,p}}{\sum_i^n \Delta t_i^{l,p}} = \frac{\sum_i^n c_i^{l,p} \times d^l}{\sum_i^n \Delta t_i^{l,p}}, \tag{S1}$$

Equation (S1) can be rewritten to Equation (S2) taking the vehicular mean speed $v_i^{l,p}$ into account:

$$V^{l,p} = \frac{\sum_i^n c_i^{l,p} \times d^l}{\sum_i^n \frac{c_i^{l,p} \times d^l}{v_i^{l,p}}} = \frac{\sum_i^n c_i^{l,p}}{\sum_i^n \frac{c_i^{l,p}}{v_i^{l,p}}}, \tag{S2}$$

Matrix of mean speed flow

Before computing average link speed flow, the data is cleaned to deal with outliers and remove local roads of minor importance, namely links frc 7. Indeed, these roads are deleted to respect the rule explained in section 2.2.1. of the article on the one hand, and on the other hand to avoid noise in the training RF model. The closest speed limit to ffs (e.g., 50 km/h for an ffs equal to 45 km/h) is imputed to just over a hundred missing speed limit data. Outliers meeting the following conditions are removed to construct the matrix of mean speed flow:

- Observed speeds higher than the speed limit increased by 20 km/h;
- Observed speeds lower than 5 km/h;
- Observed coverages equal to 0.

One matrix is built for each of the eleven regions. Several features are collected from link information provided by Mediamobile, such as link ID, ffs, frc, netclass, fow, and speed limits. Link IDs are reindexed inside each region to normalize data later. Laña et al.

designated some temporal features with a feature of cluster classes to train an RF classifier because of the daily cyclical nature of the traffic pattern [1]. Our missing data imputation method is definitely different, but inspired the way traffic conditions are reconstruct. Indeed, we built several features to take spatiotemporal relations between links into account, such as the day of the month (from 1 November to 30 November), the day type "fday" (business day, bank holiday, weekend, holiday, or weekend in holiday), the fifteen minute period of the day (from 01 to 96), and two spatial similitudes. The 15-minute period is represented by an index corresponding to the time of the day to indicate which part of the day is being considered. These two spatial similitudes are shown in Figure S3 built as follow:

1. Each of the eleven regions is divided into three subregions, except for the urban motorway, which is divided into four subregions; thus, each link is attributed a class between one and three (one and four for the urban motorway);
2. Each subregion is divided into two classes, one and two.

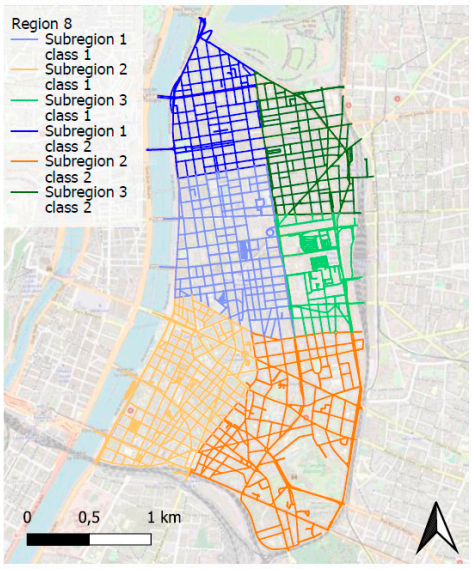


Figure S3. An instance of subregion and class for region 8.
(*background map: © OpenStreetMap*)

Finally, a matrix of these features is developed for each region with observed vehicle speeds, if any for each day, period, and link inside the region considered. Then, the mean speed flow of each link is computed as described in Equation (S2) for each period of fifteen minutes on the whole November month, as illustrated in Table S3.

Table S3. Extract of matrix of mean speed flow in region 0.

Link ID	Link ID reindexed	Observation	length (km)	coverage	ffs (km/h)	Speed limit (km/h)	frc	netclass	fow	Subregion	Class	day	fday	period	$c_i \times d_i$	$\frac{c_i \times d_i}{v_i}$	Mean speed flow rounded
1250000971474	0	NA	0.06134	NA	50	50	1	1	10	2	2	1	4	1	NA	NA	NA
1250000971474	0	1	0.06134	1	50	50	1	1	10	2	2	1	4	2	0.06134	0.00095844	56
1250000971474	0	2	0.06134	1	50	50	1	1	10	2	2	1	4	2	0.06134	0.00115736	
1250000971474	0	3	0.06134	1	50	50	1	1	10	2	2	1	4	2	0.06134	0.00115736	
1250000971474	0	4	0.06134	1	50	50	1	1	10	2	2	1	4	3	0.06134	0.00191688	33
1250000971474	0	5	0.06134	1	50	50	1	1	10	2	2	1	4	3	0.06134	0.00175257	
1250000971474	0	6	0.06134	1	50	50	1	1	10	2	2	1	4	4	0.06134	0.00097365	54
1250000971474	0	7	0.06134	1	50	50	1	1	10	2	2	1	4	4	0.06134	0.00130511	
…	…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	
1250000971474	0	i	0.06134	1	50	50	1	1	10	2	2	1	4	93	0.06134	0.00142651	47
1250000971474	0	i+1	0.06134	1	50	50	1	1	10	2	2	1	4	93	0.06134	0.00120275	
1250000971474	0	NA	0.06134	NA	50	50	1	1	10	2	2	1	4	94	NA	NA	NA
1250000971474	0	i+2	0.06134	1	50	50	1	1	10	2	2	1	4	95	0.06134	0.0014961	44
1250000971474	0	i+3	0.06134	1	50	50	1	1	10	2	2	1	4	95	0.06134	0.0014961	
1250000971474	0	i+4	0.06134	1	50	50	1	1	10	2	2	1	4	95	0.06134	0.00115736	
1250000971474	0	NA	0.06134	NA	50	50	1	1	10	2	2	1	4	96	NA	NA	NA

Point S2

RF Machine Learning Workflow

The workflow to generate efficient RF models is defined by the implementation of the following steps for each region:

1. The matrix of the mean speed flow is built, so the values of features ranged over different scales and must be normalized using reindexed Link ID as the main scale in Equation S3 below:

$$X_{norm,i} = \frac{(X_i - X_{min,i})}{(X_{max,i} - X_{min,i})} \times N_{links,i}, \quad \text{(S3)}$$

   where X_i is the feature in region i at its original scale, $X_{min,i}$ and $X_{max,i}$ are respectively the minimum and the maximum of feature X_i, $X_{norm,i}$ is the normalized feature in region i at reindexed Link ID scale, and N_{links} is the number of links in region i;
2. The entire dataset of each region is split into two subsets: a train dataset containing 75% of the whole dataset and a test dataset covering 25%. This splitting process is carried out to identify the best depth of trees and avoid overfitting and underfitting;
3. The RF model is trained on the training set with two hundred decision trees and different depths;
4. The trained RF model is then used to infer seen data on the training dataset and unseen data on the test dataset;
5. The inferred data are evaluated with RMSE and R^2 for each depth parameter;
6. The RF model inferring the best performance for both subsets is selected;
7. Finally, the RF model is trained on the entire dataset with two hundred decision trees and the best depth parameter for each region and used to infer the missing mean speed flow.

Figure S4 illustrates the workflow to construct RF models. Figure S5 proposes a heat map of mean speed flow for a sample of Link ID in region 0 on November 6, 2017 for observed and missing mean speed flow in (a) and observed and deduced mean speed flow by RF in (b). Table S4 shows the results of RMSE and R^2 for RF model on training and testing dataset with the best hyperparameters.

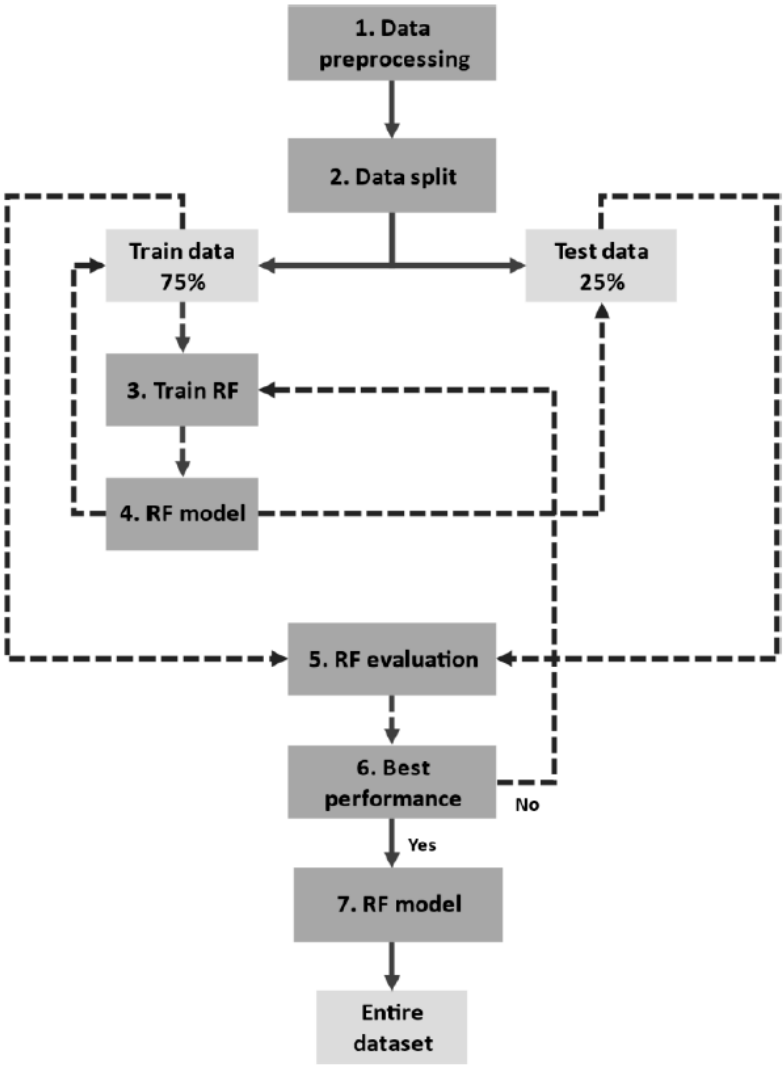


Figure S4. Workflow of RF training.

Table S4. Results of RMSE and R^2 for RF model on training and testing dataset.

Region	R^2		RMSE	
	Train	Test	Train	Test
0	0,80	0,68	6,74	8,58
1	0,73	0,56	7,07	9,00
2	0,84	0,71	6,67	9,15
3	0,74	0,59	7,67	9,72
4	0,84	0,72	6,89	9,06
5	0,78	0,63	7,06	9,27
6	0,77	0,64	7,37	9,25
7	0,78	0,62	6,76	8,83
8	0,68	0,53	6,60	8,04
9	0,85	0,68	5,79	8,53
Urban motorway	0,95	0,86	4,14	7,16

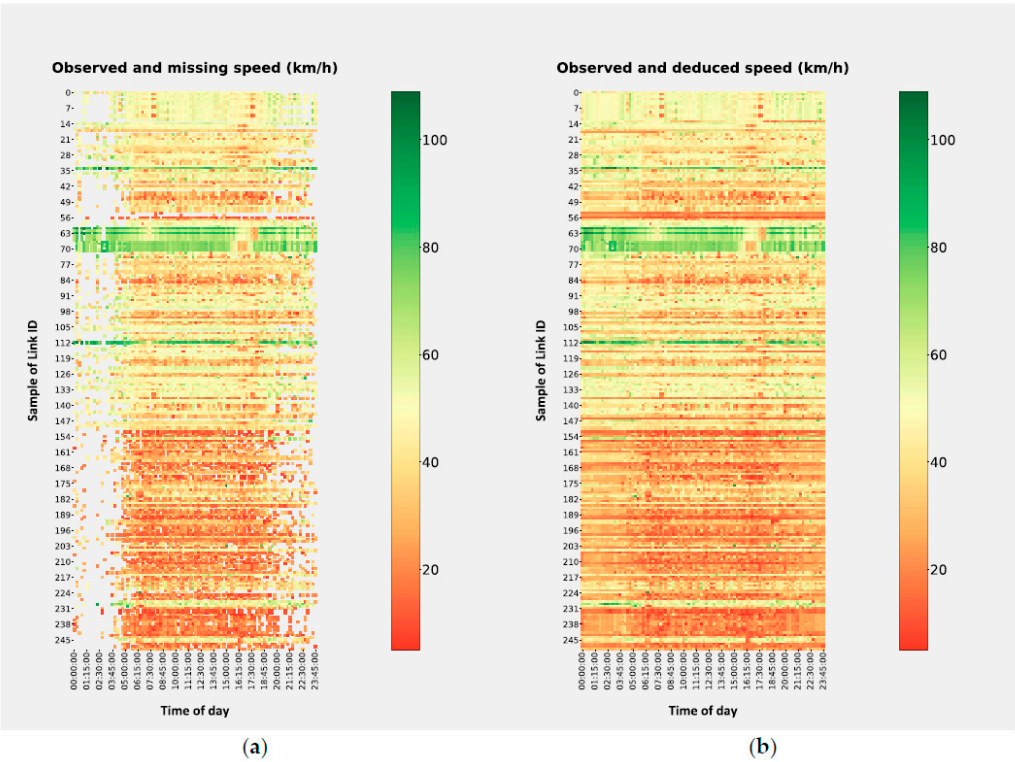


Figure S5. Heat map of mean speed flow for a sample of Link ID in region 0 on November 6, 2017: (**a**) Observed and missing mean speed flow (**b**) Observed and deduced mean speed flow by RF.

Point S3

Algorithm S1. Trip enhancement

```
Define subset trips = both continuous temporal and discontinuous spatial trips
for trip ϵ subset trips do:                                   Loops to each trip in the subset
    Extract start and end nodes of each edge traveled
    for edge e and e + 1 traveled ϵ trip do:                  Loops to each consecutive edge traveled in the selected trip
        if the end node of edge e ≠ the start node of edge e + 1 then:
            Compute SP with Dijkstra algorithm
            if SP > threshold then:
                Stock Vehicle ID of the trip
return all discontinuous trip enhanced
Remove trips with SP > threshold
```

Point S4

This Point S4 shows the same results as section 3.1.2. in the article for FC and every other pollutant studied (NOx, PM_{10} and Multipollutant). Figures S6 and S7 displays respectively their DR_{SuP} distribution and ΔTT distribution. Tables S4, S5, S6 and S7 presents their summary statistics.

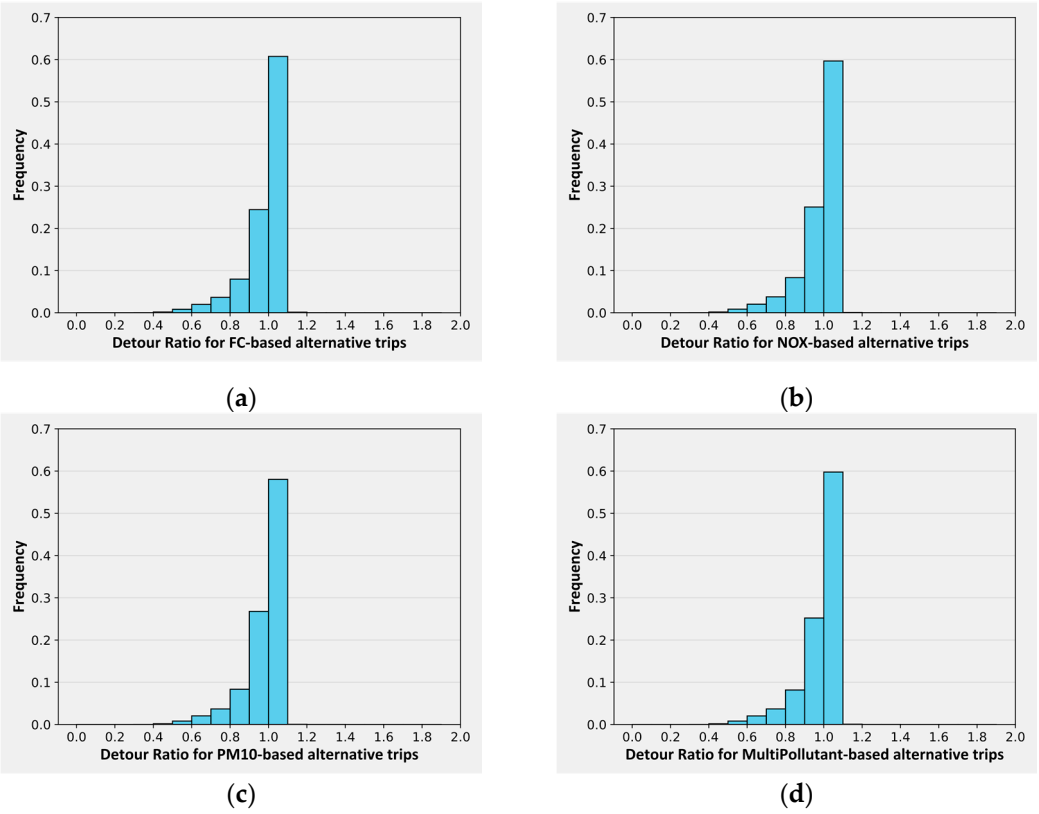


Figure S6. DR_{SuP} distribution: (**a**) FC-based alternative trips (**b**) NOx-based alternative trips (**c**) PM_{10}-based alternative trips (**d**) Multipollutant-based alternative trips.

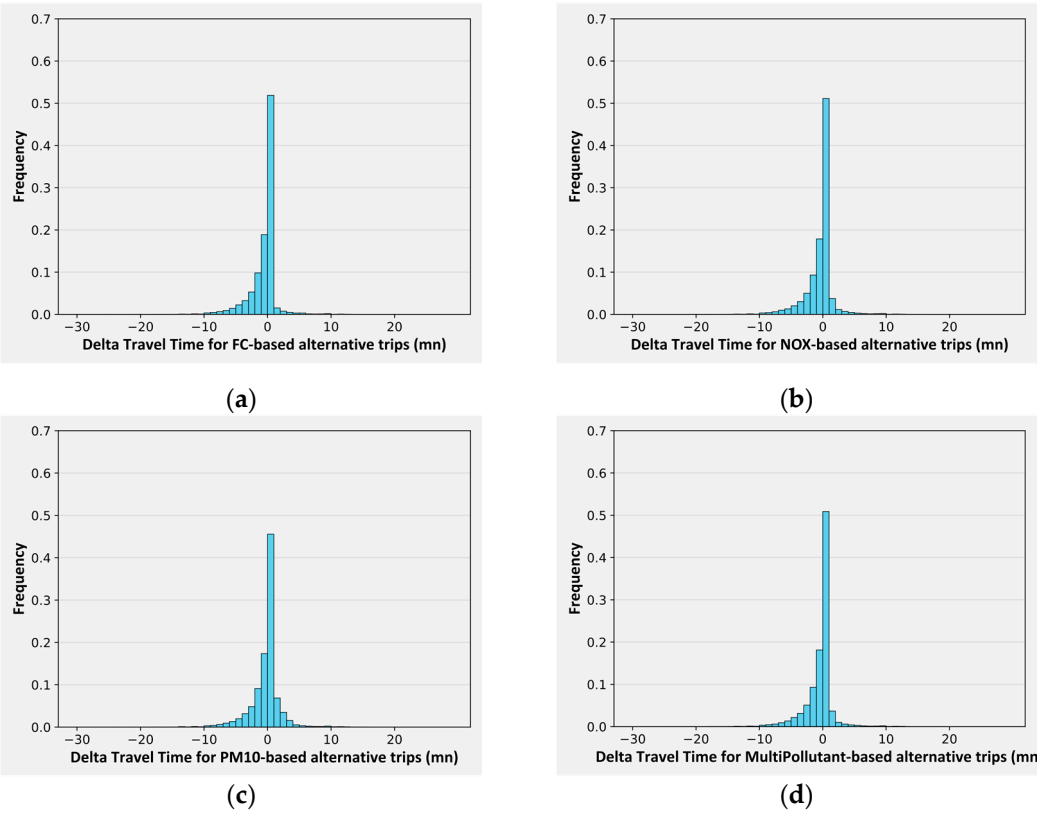


Figure S7. ΔTT distribution: (**a**) FC-based alternative trips (**b**) NOx-based alternative trips (**c**) PM_{10}-based alternative trips (**d**) Multipollutant-based alternative trips.

Table S5. Summary statistics of FC-based alternative trips.

Statistics	TD (km)	TT (min)	TS (km/h)	TE FC (litres)	DR_{SuP} (no unit)	ΔTT (min)
Mean	7.99	11.08	50.42	0.19	0.96	-0.70
Standard deviation	3.55	5.03	17.66	0.075	0.09	2.24
Minimum	0.68	2.46	6	0.03	0.37	-25.73
25th percentile	5.15	7.26	34.47	0.13	0.97	-1.00
50th percentile	7.63	10.02	50.43	0.18	1	0
75th percentile	11.24	13.46	67.65	0.26	1	0
Maximum	18.55	43.94	86.13	0.44	1.35	13.40

Table S6. Summary statistics of NOx-based alternative trips.

Statistics	TD (km)	TT (min)	TS (km/h)	TE NOx (g)	DR_{SuP} (no unit)	ΔTT (min)
Mean	7.98	11.24	49.86	4.06	0.96	-0.54
Standard deviation	3.55	5.12	17.32	1.65	0.09	2.29
Minimum	0.68	2.46	6	0.56	0.36	-23.46
25th percentile	5.12	7.35	34.13	2.78	0.96	-0.91
50th percentile	7.63	10.10	50.19	3.86	1	0
75th percentile	11.19	13.60	66.78	5.59	1.00	0.05
Maximum	17.38	43.94	86.13	9.22	1.35	15.98

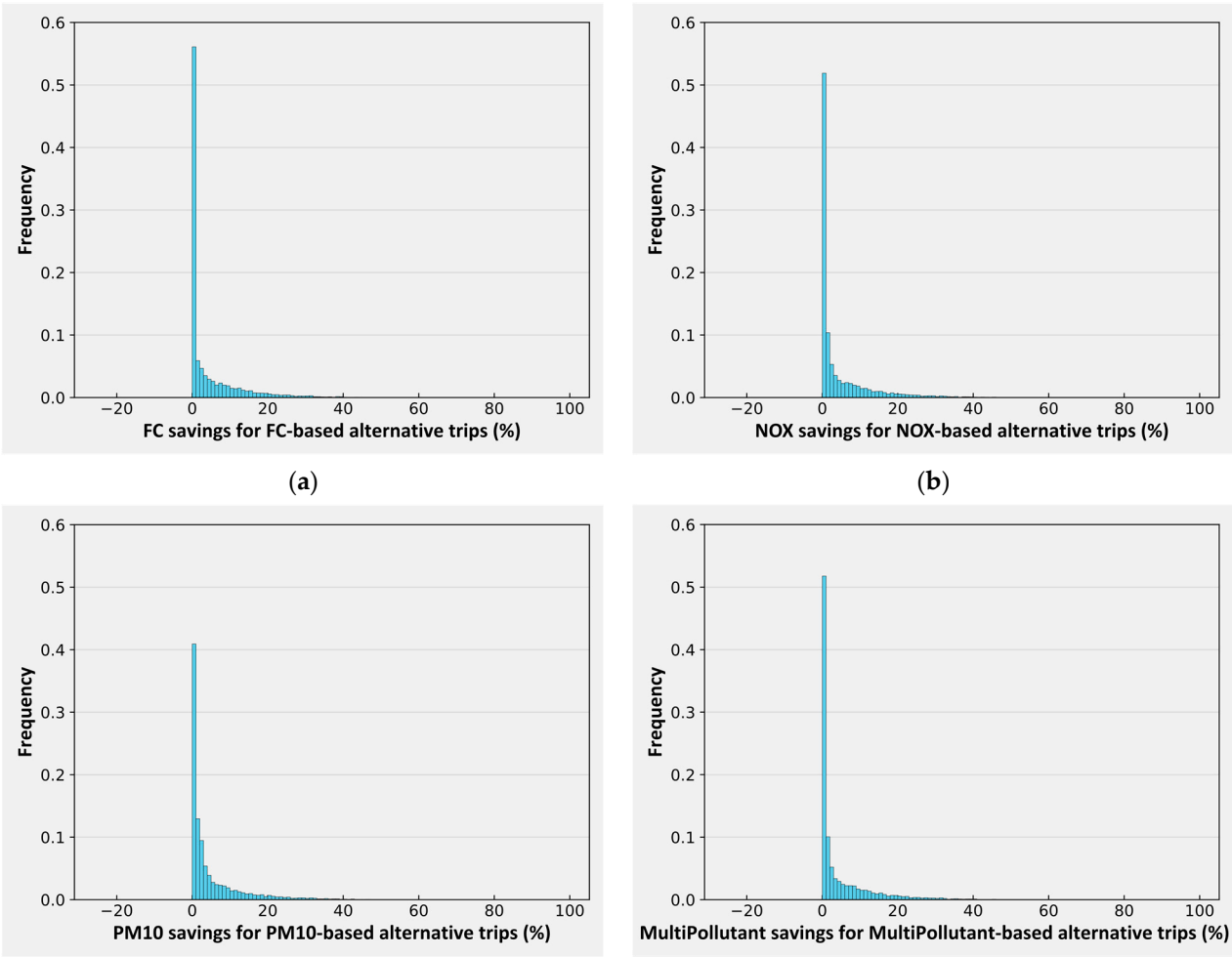
Table S7. Summary statistics of PM_{10}-based alternative trips.

Statistics	TD (km)	TT (min)	TS (km/h)	TE PM_{10} (g)	DR_{SuP} (no unit)	ΔTT (min)
Mean	7.98	11.40	49.05	0.09	0.96	-0.38
Standard deviation	3.55	5.14	16.58	0.03	0.09	2.39
Minimum	0.68	2.46	6	0.01	0.37	-23.46
25th percentile	5.13	7.45	34.19	0.06	0.96	-0.87
50th percentile	7.64	10.30	49.46	0.08	1	0
75th percentile	11.19	13.81	64.43	0.12	1.00	0.29
Maximum	17.38	43.94	86.13	0.19	1.35	15.49

Table S8. Summary statistics of Multipollutant-based alternative trips.

Statistics	TD (km)	TT (min)	TS (km/h)	TE Multi (no unit)	DR_{SuP} (no unit)	ΔTT (min)
Mean	7.98	11.21	49.89	4.38e-3	0.96	-0.57
Standard deviation	3.55	5.08	17.25	1.77e-3	0.09	2.28
Minimum	0.68	2.46	6	0.63e-3	0.37	-23.46
25th percentile	5.13	7.34	34.25	3e-3	0.96	-0.93
50th percentile	7.64	10.09	50.18	4.16e-3	1	0
75th percentile	11.20	13.58	66.77	6.02e-3	1.00	0.05
Maximum	17.38	43.94	86.13	10.07e-3	1.35	14.54

Point S5

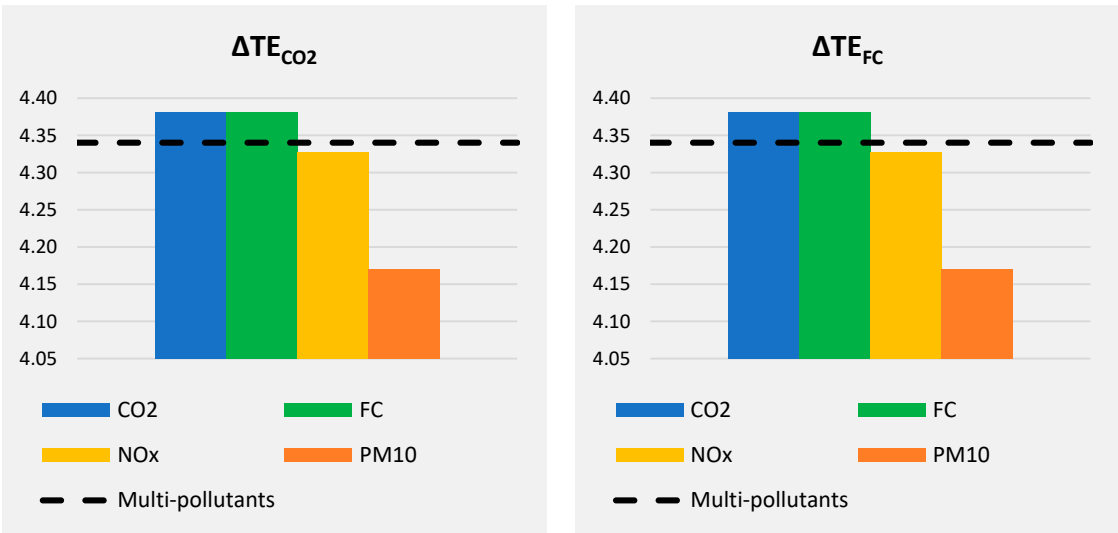


(**a**) (**b**)

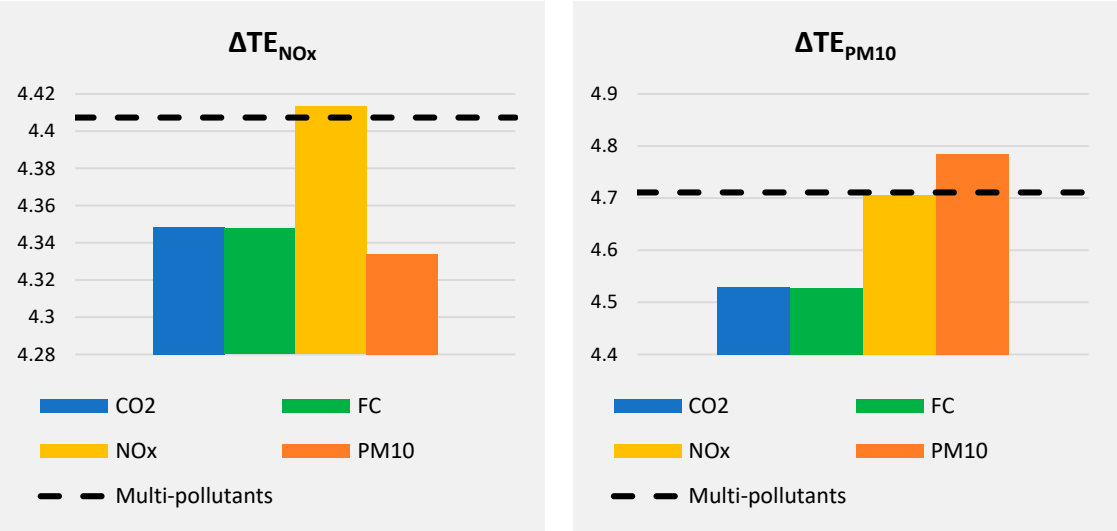
(**c**) (**d**)

Figure S8. Emission saving distribution: (**a**) FC for FC-based alternative trips (**b**) NOx for NOx-based alternative trips, (**c**) PM_{10} for PM_{10}-based alternative trips, (**d**) Multipollutant for Multipollutant-based alternative trips.

Point S6



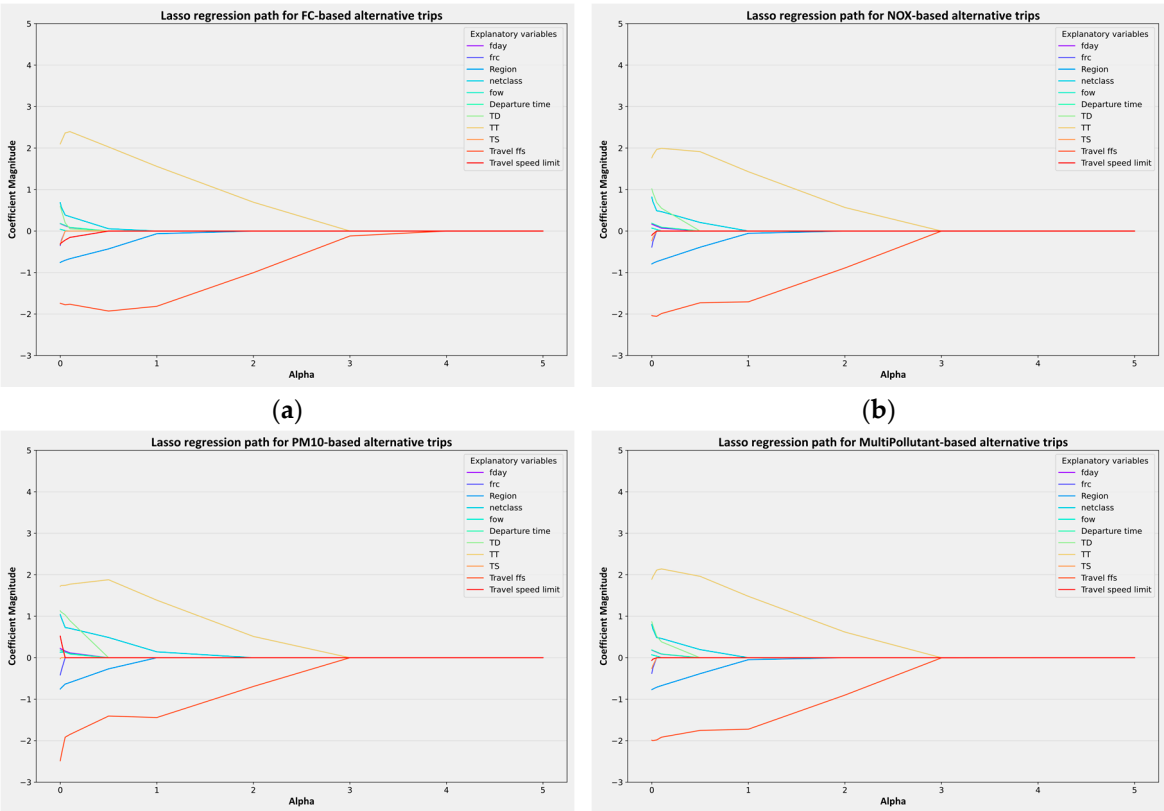
(**a**) (**b**)



(**c**) (**d**)

Figure S9. Average emission savings per trip for: (**a**) CO_2 for CO_2-based alternative trips (**b**) FC for FC-based alternative trips (**c**) NOx for NOx-based alternative trips (**d**) PM_{10} for PM_{10}-based alternative trips.

Point S7



(**a**) (**b**)

(**c**) (**d**)

Figure S10. LASSO regression results by Alpha coefficient for: (**a**) FC-based alternative trips, (**b**) NOx-based alternative trips, (**c**) PM_{10}-based alternative trips, (**d**) Multipollutant-based alternative trips.

Reference

1. Laña, I.; Olabarrieta, I. (Iñaki); Vélez, M.; Del Ser, J. On the Imputation of Missing Data for Road Traffic Forecasting: New Insights and Novel Techniques. *Transp. Res. Part C Emerg. Technol.* **2018**, *90*, 18–33, doi:10.1016/j.trc.2018.02.021.