Deliverables for the Project: Facility Location Choice Simulation Tool (FaLC)

Transport Simulation Module: Speed Regression

Deliverables for the Project:

Facility Location Choice Simulation Tool (FaLC)

Transport Simulation Module: Speed Regression

Georgios Sarlas
Kay W. Axhausen

January 2014

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Table of contents

Tables

Figures

1 Introduction – Objective

The objective of the current module is to provide a direct and simplified way of estimating velocities for each link of the network of interest. In order to accomplish this, regression modeling is the chosen approach. The main advantage of that choice, in comparison to more advanced and complex demand models (eg four-step travel demand modelling approach, agent-based approaches), is that it can provide a structural equation of the speeds in association with the characteristics that affect the demand for travel but in a significantly less cumbersome way than the aforementioned approaches. The attributes that are used in the regression correspond to only aggregated values where no personal data information can be traced back. More specifically, socio-demographic attributes along with attributes that represent the characteristics of the network are taken into account for that purpose.

The objective of the current work is to provide a coherent transport simulation module (regression model) that can be employed within an overall framework of a land-use and transport interaction tool in order to assist a facility location choice analysis.

2 Data preparation and analysis

In the following section the preparation and the analysis of the required data for the regression model is presented. The main challenge of this particular project lies on the successful combination of data coming from various sources, thereupon the data preparation constitutes an important part of the project and as such is presented analytically. In summary, data can be classified into two distinct categories; network and socio-demographic data. The data preparation, analysis and the model estimation are conducted in R, unless stated otherwise.

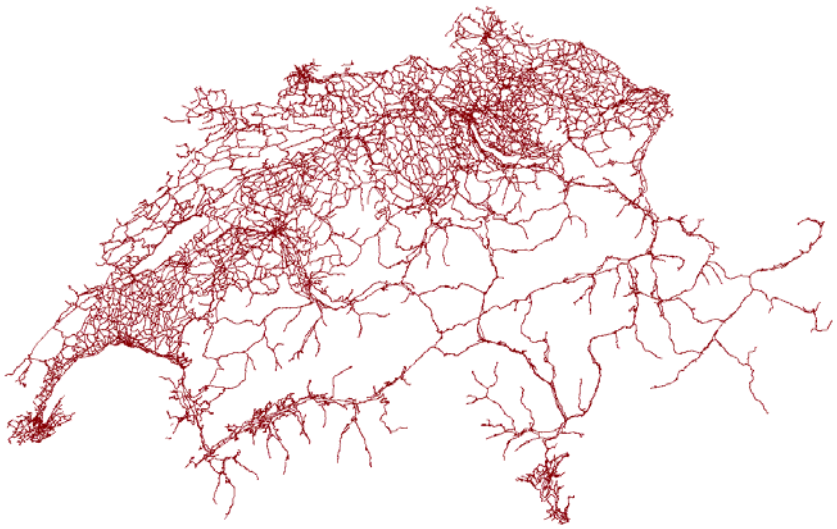
2.1 Network data

2.1.1 Road network

The network that is employed for the current project is an open-street map (OSM) network. More specifically, a sample of the freely available full network of Switzerland is used where the links of interest were determined and included in the network of study. In summary, the resulted network of study consists of 50.293 links, out of which 36.366 are bidirectional. A visual representation of the network is presented below.

The spatial location and the shape of each link are included in the network file, which is in a shapefile format. Apart from these attributes, also the following attributes are included per link: a unique identifier, name, type, maximum speed, if the link is bidirectional, and if it is bridge. Based on the nature of the OSM, where users are updating and editing the attributes of the network, the problem of missing values arises. Ideally, max-speed would be the most appropriate value to allow a consistent classification of the links for the speed regression. However, almost 2/3 of the links don't have a reported max-speed value. Nevertheless, a categorization of the links per type exists which will be utilized. In the network of study, ten different type of links can be found which are presented in the following table along with the number of links of each type.

Figure 1: The network of study (open-street map network)



In order to assess the quality of the reported classification of the links, the maximum speed per link type is plotted. Consequently, the problem of the missing values comes to the surface since as it can be seen in Fig.2, the range of max-speed values per type is pretty wide, while the average max-speed per type is close to zero for most of the link types. Due to this fact, it is considered essential to plot the max-speed per type, excluding though the links that do not have a reported maximum speed in order to obtain an accurate idea of the consistency of the classification of the links based on their reported max-speed. As it can be seen in Fig.3, the deviation of the max-speed per type is much lower than before and also the respective average values is much more reasonable and according to our expectations. Based on the above, it can be concluded that the reported classification of the links of the network seems to be accomplished in a consistent way, with the exception of very few links (represented as circles in the figure).

Table 1: Number of links per type in the network of study

Type of link	Number of links
motorway	3787
motorway_link	3107
primary	10780
primary_link	591
secondary	12137
secondary_link	52
tertiary	18333
tertiary_link	31
trunk	894
trunk_link	581
Total	50293

Figure 2: Boxplot of maximum speed per reported type of link

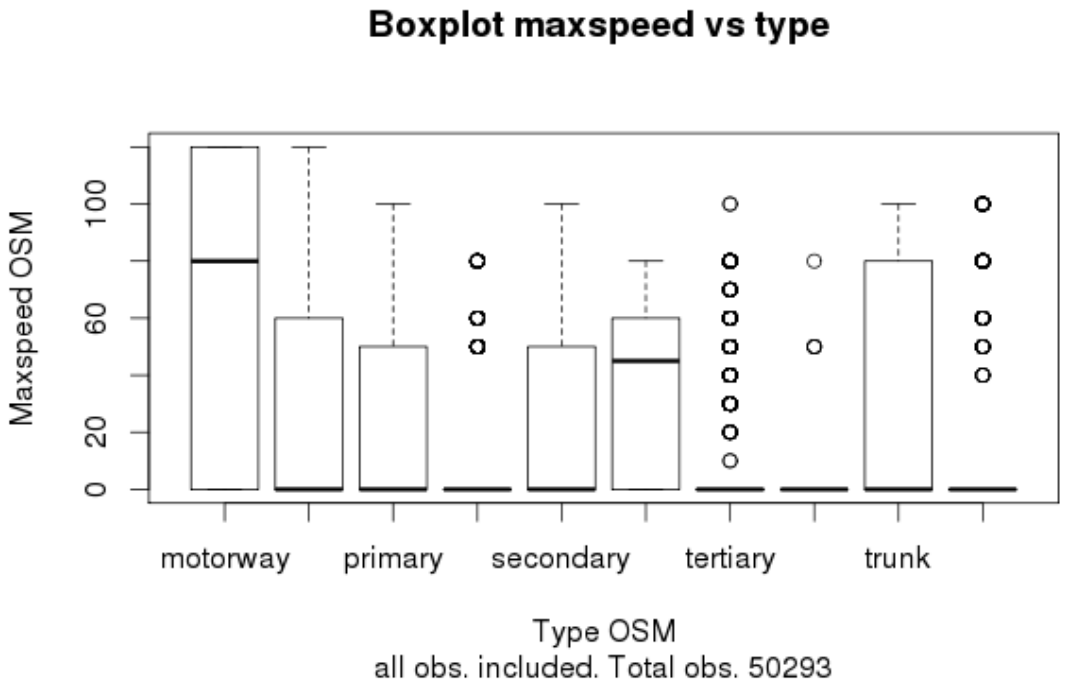
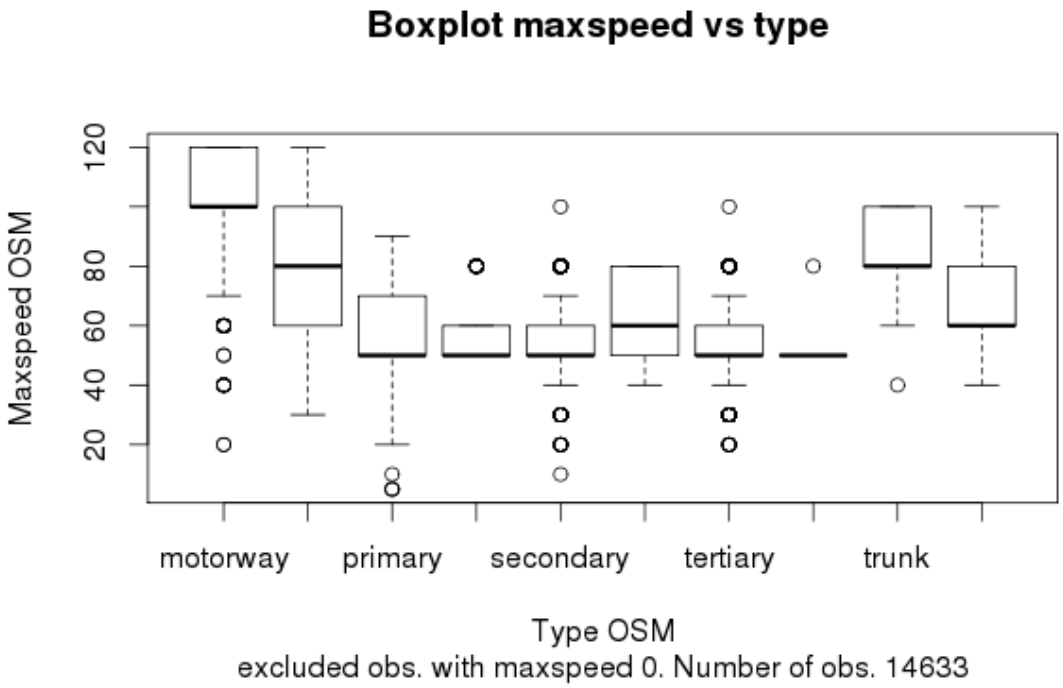


Figure 3: Boxplot of maximum speed per reported type of link, excluding links with zero maximum speed



It should be noted that all the spatial data used are, or transformed accordingly if needed, in the Swiss reference system CH1903.

2.1.2 Public transport network

Public transport network constitutes an important part of the transportation system and its impact on road speeds cannot be neglected, especially in the urban areas. In order to account properly for that impact, some attributes capturing the density of public transport network need to be included. The most appropriate representative attribute is considered to be the density of public transport stops within a given area. These data are extracted from the Swiss National Transport model (version: year 2007).

2.2 Socio-demographic data

The socio-demographic data that are used in this project are available from the Swiss Federal Statistical Office (BFS: Bundesamt für Statistik). More specifically, the socio-demographic

data of interest are the population and the employment positions for the whole area of Switzerland, aggregated per hectare. The population data from the year 2011, taken from the “Statistics of Population and Households 2011” ("Statistik der Bevölkerung und der Haushalte 2011", date of version: 30 August 2012), while the employment data are taken from the “Federal Business Census” of 2008 (Eidgenössische Betriebszählung 2008, date of version: 29 March 2008).

3 Model estimation

In this section, the process followed to estimate the speed regression model, along with the required calculation of the different explanatory variables that its inclusion in the specification of the model is tested, is presented. It is reminded that the purpose of the regression model is to be applied to predict the speeds on the links of the entire network of study. The average daily speed for a typical weekday is the dependent variable of interest. The regression yields two speed components; first, the average road speed per road type, is a non-spatial quantity. Spatial variation is added to the link speed estimates in the second component via the spatially resolved explanatory variables. Spatially resolved road and public transport network densities represent the effect of road supply on speed. Spatial data on population and employment is taken to be indicative of the intensity of local activities, reflecting travel demand locally Hackney et al. (2007).

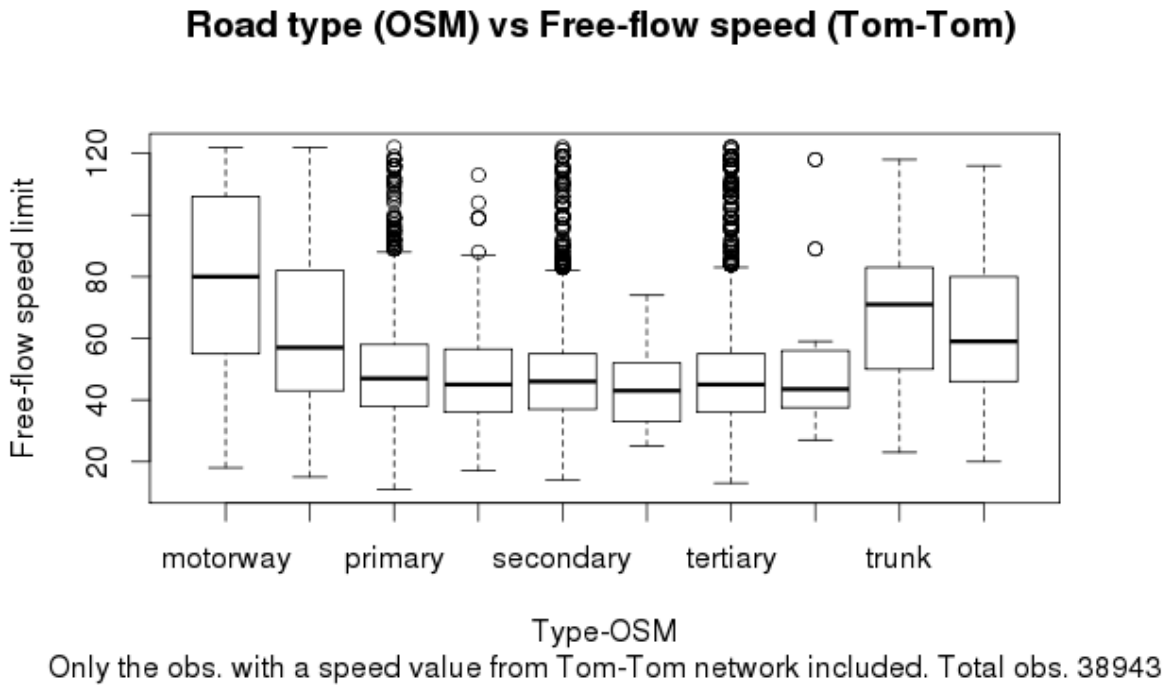
3.1 Actual link speeds

In order to proceed to the estimation of the speed regression model, a set of actual speeds is prerequisite to be used as target values for the regression. The target values are available in a different network (commercially provided by Tom-tom), where the average daily speed for weekdays is estimated based on GPS measurements, on a link level. In order to overcome the limitation of different networks’ definition, the employed network is spatially joined with the Tom-tom network in order to transfer the actual speed values of each link to the OSM network. ArcGIS software is used for that purpose and in particular the function of “Spatial Joint” under the “Spatial Analyst” menu. Approximate 22% of the links included in the network of study, are not spatially joined to a link from the Tom-tom network, or are joined to a link that does not have an estimated average speed.

In the following figure, the free flow speed of the spatially joined links of the OSM network (transferred from Tom-tom network) is plotted against the type of the links. Attempting a

comparison with the plot in Fig.3 and assuming that the free flow speed variable is the closest equivalent to the maximum speed variable of OSM network, it can be seen that the average free flow speed is a bit lower than in Fig.3, and also the deviation from the average values is larger than before. This difference can be explained mainly by the different network definitions and also up to an extent can be attributed to wrongly spatially joined links. Additionally, it cannot be overlooked the nature of the OSM network where the variables, and thus the maximum speed, are registered by individual users, who might perceive as higher the actual speed than it really is, and thus report it mistakenly. In summary, the same trend as before can be seen regarding the average speeds, and thus it can be concluded that the outcome of the spatial join process is not perfectly accurate since links' definition varies between the two networks, however it is considered to be sufficiently accurate for the purposes of this project.

Figure 4: Boxplot of free-flow speed per reported type of link after the spatial join of the networks



3.2 Explanatory variables

As mentioned before, the second component consists of spatially resolved explanatory variables. Spatially resolved road and public transport network densities represent the effect of road supply on speed. The corresponding road (line) densities are calculated in ArcGIS and the results are stored in cells (pixels) of 100 x 100 meters dimensions to coincide with the

cells from the socio-demographic data (hectare data). The results are expressed in meters (length) per square kilometers. In order to calculate the density of the OSM network, the full OSM network of Switzerland was fetched for that reason (date: 30 September 2013). The line density is calculated over different radii R of 100 meters, 500 meters, 1 kilometer, and 5 kilometers. Apart from the total road network density, the density per link type is calculated as well to examine the impact of specific road types' density on the speed, over radii of 500 meters and 1 kilometer. The same density measurements are calculated also for the Tom-tom network and the Navteq network (commercial navigational network) to compare the results and assure that the calculation is made in a correct way. The results are similar. A visual representation of the road density of OSM network over 500 meters radius, as well the density of residential streets over the same radius is presented below.

Figure 5: Road density of OSM network over 500 meters radius

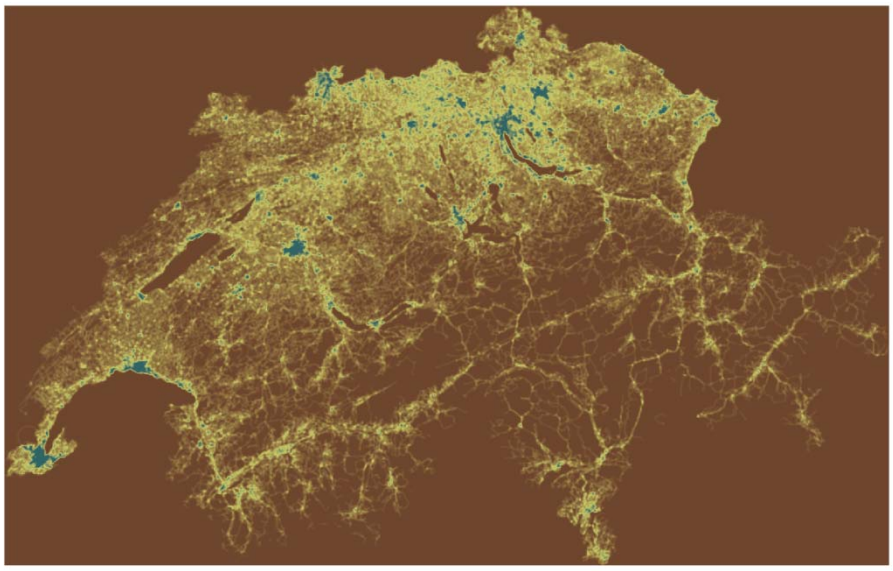


Figure 6: Residential road density of OSM network over 500 meters radius

As it can be seen in Fig.6, the residential road density highlights the main urban and dense areas of Switzerland.

In the case of the public transport network and its density, it is considered more appropriate to calculate the density of the number of stop points as more representative measure. As such, the point density is calculated using ArcGIS software over the same radii as the previous densities. In the following plot, the calculated density of public transport stop points can be seen over a radius of 500 meters. Each point is taken into account multiple times equal to the number of lines that serve each particular stop. A visual representation of the public transport stop points density is presented below.

In the same way, the spatial explanatory variables of population and employment densities are calculated. More specifically, the variables of the total population and the full-time equivalent employment positions per hectare are created by aggregating the corresponding values. In this particular case, both variables' densities are also kernel weighted to capture their diminishing impact on the speeds over distance. The same radii as before are used for the calculation and they are expressed in population/ employment positions per square kilometer. In the plots that follow, the difference between the normal and the kernel density calculation is exhibited by focusing in the wider area of Zurich. As it can be seen, the kernel weighted densities are smoother in space.

Figure 7: PuT stop points densities over 500 meters radius

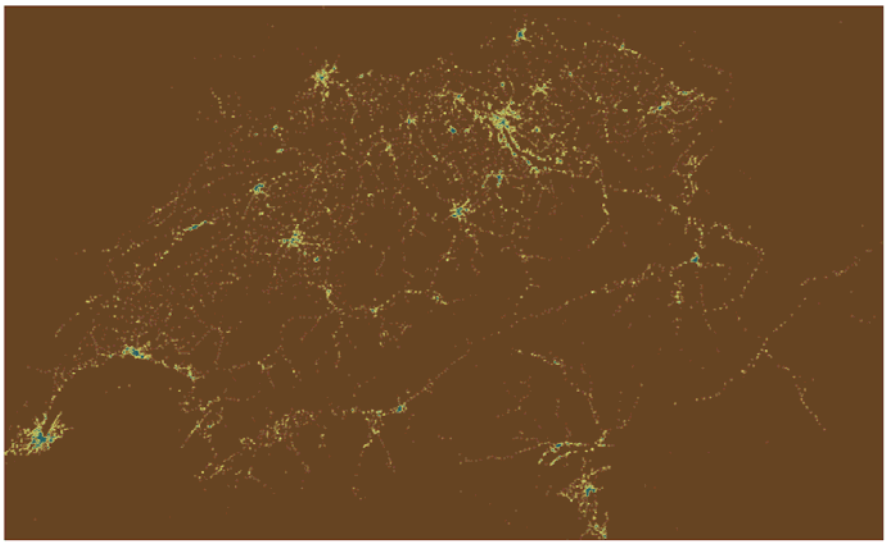


Figure 8: Normal density of population over 1 kilometer radius

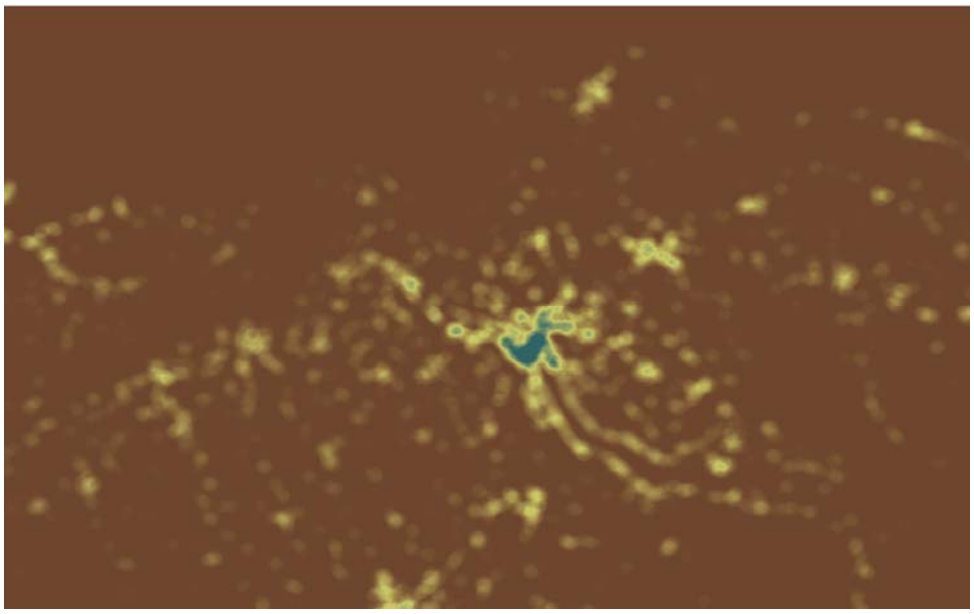
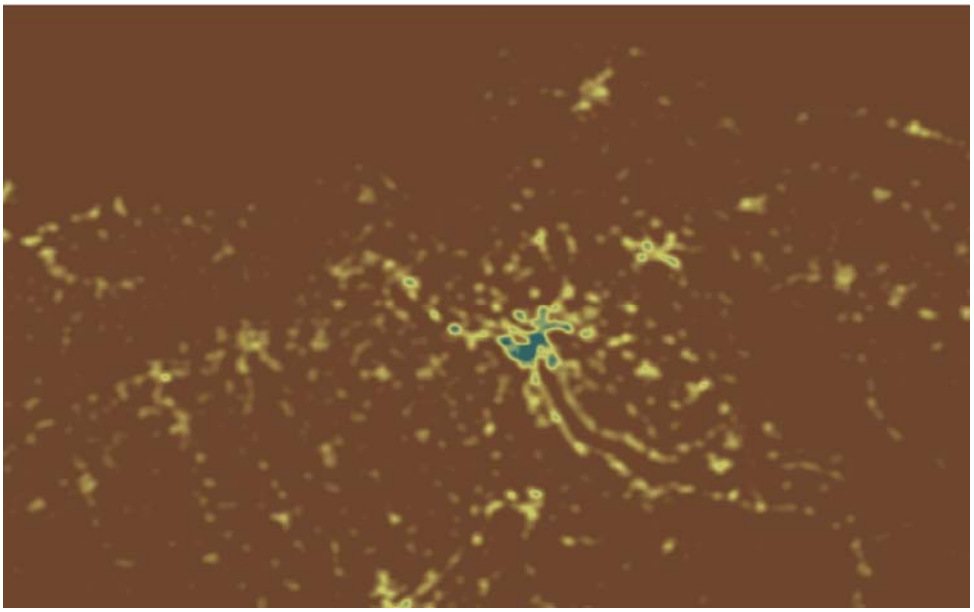


Figure 9: Kernel weighted density of population over 1 kilometer radius



The spatial explanatory variables calculated so far need to be associated to the links of the network. Thereupon, each link of the network associate with the hectare (cell) values of each spatial variable, closest to the upstream endpoint of the link.

Apart from the above spatially explanatory variables, the average gradient (slope) of each link is an important variable that its inclusion in the model should be tested. A priori, it is expected that uphill results in reductions on the speeds, especially in the case of lower classified roads (e.g. secondary streets). A terrain model with a resolution grid of 25 x 25 meters is used in order to infer the average gradient of each link of the network by intersecting the terrain model with the nodes of the network.

Another important variable that should be tested, corresponds to the decrease of the speed as a result of the link being an on/off ramp to highway. In order to accomplish that, the classification of the Tom-tom network is utilized respectively since there is no such information in the OSM network. The results from the spatial join process that preceded, are used to accomplish that and identify the ramps of the network. In addition, the ramp density is calculated as well as line density in the same way as before over 500 meters and 1 kilometer radii.

A list of the available variables is shown in the table below.

Table 2: Overview of the available attributes for the regression model.

Source		List of variables	Values	
OSM		Type	factor	1-10
		maxspeed	numerical	0-120
		Oneway	dummy	
		Bridge	dummy	
Terrain model		average gradient per link (percent and angle)	numerical	
Tom-tom		Type	factor	0,1,99
		Ramp	dummy	
		speed limit	numerical	
		speed free flow	numerical	
		avg speed weekday	numerical	
Employment		Radius of 0.5km	normal & kernel densities	
		Radius of 1km		
		Radius of 2km		
		Radius of 5km		
Population		Radius of 0.5km	normal & kernel densities	
		Radius of 1km		
		Radius of 2km		
		Radius of 5km		
PuT stops (National model 2007)		Radius of 200m	normal & kernel density	
		Radius of 0.5km		
		Radius of 1km		
		Radius of 2km		
		Radius of 5km		
Network Densities	OSM	Radius of 100m	Line densities	
		Radius of 0.5km		
		Radius of 1km		
		Radius of 5km		
		per type, radius of 0.5km and 1km (for all types)		
	Tom-tom	Radius of 100m		
		Radius of 0.5km		
		Radius of 1km		
		Radius of 5km		
	Navteq	Radius of 100m		
		Radius of 0.5km		
		Radius of 1km		
		Radius of 5km		
		per type, radius of 0.5km and 1km (for all types)		
On/Off Ramps		Radius of 0.5km and 1km	Line densities	
Population and employment		all from BFS per hectar	numerical	

3.3 Specification and estimation of model

The next step after the preparation of the set of variables is to proceed to the specification of the regression model and the estimation of its corresponding parameters. The estimation of the regression model is conducted in R, by means of ordinary least squares estimation (OLS). Different specifications of the model are checked based on the set of the available variables, where variables are added gradually and their explanatory power is evaluated in terms of magnitude, sign, substitution ratios, statistical significance (t-test), and goodness of fit measures (adjusted coefficient of determination, Akaike Information Criterion (AIC)). Based on the above, the specification of the model is chosen in line with having a parsimonious model. The estimated values of its regressors are presented in table 3. Based on the results from the different tested specifications, and more specifically due to the fact that some of the regressors of the different types of link (average speed) have almost identical values, some link types are merged together (eg motorway and motorway link), resulting to a reduction in the number of variables in a coherent and consistent way (six types of links instead of 10).

Table 3: Estimated parameters of the speed regression model

Coefficients	Estimate	Std. Error	t- value	Pr(> |t|)
Type: motorway	8.50E+01	3.74E-01	227.054	< 2E-16
Type: motorway link	7.33E+01	4.41E-01	166.052	< 2E-16
Type: primary	5.73E+01	2.46E-01	232.771	< 2E-16
Type: secondary	5.57E+01	2.39E-01	233.551	< 2E-16
Type: tertiary	5.39E+01	2.12E-01	253.659	< 2E-16
Type: trunk	7.13E+01	5.52E-01	129.258	< 2E-16
Ramp (dummy)	-9.30E+00	4.03E-01	-23.095	< 2E-16
PuT stops 0.5km	-4.36E-01	2.75E-02	-15.859	< 2E-16
Employm.Normal 0.5km	-2.25E-04	3.13E-05	-7.189	6.64E-13
Population kernel 0.5km	-6.45E-04	4.47E-05	-14.433	< 2E-16
Osm residential density 0.5km	-5.59E-01	1.76E-02	-31.8	< 2E-16
Osm line-density 0.5km	-9.68E-02	9.25E-03	-10.467	< 2E-16
Type secondary: gradient	-6.09E-02	4.89E-02	-1.244	0.21338
Type tertiary: gradient	-9.55E-02	3.58E-02	-2.668	0.00764
Type trunk: gradient	-1.67E-01	1.40E-01	1.40E-01	0.23242

Number of observations: 38943
Mult. Adjusted R-square: 0.9099
AIC: 32434.8

The spatial explanatory variables that are included in the model reflect that the impact of the socio-demographic and network data is localized. More specifically, the variables of densities over a radius of 500 meters are found to give the more reasonable and statistically significant results. Population is taken into account as kernel weighted density, exhibiting a diminishing impact on the link speeds over the distance while the full-time equivalent employment positions and public transport stops are taken into account as normal densities (point densities). Regarding the road densities per type of links, it is found that the residential roads density variable should be included in the model. A finding which is in accordance to the fact that residential roads dominate over the rest road types (about 1/5 of roads are classified as residential ones), and thus the use of their density as a stand-alone variable can be considered justifiable. The inclusion of ramps density variable is checked as well but it is found not to have the expected impact on the regression (opposite sign), and thus it is excluded.

Special attention is paid on the inclusion of the gradient of the links in the regression model. Gradient as a stand-alone variable results to regressor values with opposite sign, compared to the expectation. As a consequence, gradient of the links is taken into account as an interaction with the link type, though only for the links with lower classification (secondary, tertiary, and trunk links).

It is useful to put into perspective the range of spatially resolved explanatory variables values to comprehend their impact on the speed regression model.

Table 4: Range of values of the non-dummy explanatory variables

Variable	Range of values
Density of PuT stops in 0.5 km radius	0-33.1 stops/ sqr.km
Density of employment positions in 0.5 km radius	0-46564 employm. / sqr.km
Kernel weighted density of population in 0.5 km radius	0-27428 residents /sqr.km
Density of residential links (OSM) in 0.5 km radius	0-33.5 m/ sqr.km
Density of all links (OSM) in 0.5 km radius	0-97.8 m/ sqr.km
average link gradient (%)	[-8,8]

The correlation of the non-dummy included variables is calculated to measure the linear association between them. As it can be seen in table 5, correlation values lies between 0.31 to 0.66, indicating a positive correlation. The magnitude and the sign of the correlation values comes as a consequence of the fact that these variables are reflective of the urban density. However, it is considered that each variable has descriptive power that cannot be overlooked and justify the omission of the variable.

Table 5: Correlation between the non-dummy explanatory variables

	Employm.	PuT stops	Population	Links-density	Residential density
Employm.		0.626	0.481	0.525	0.313
PuT stops	0.626		0.650	0.662	0.500
Population	0.481	0.650		0.683	0.623
Links-density	0.525	0.662	0.683		0.643
Residential density	0.313	0.500	0.623	0.643	

3.3.1 Residuals analysis

An analysis of the residuals of the estimated regression model is conducted in order to evaluate the model and comprehend its weaknesses and strengths. In Fig.11 the spatial distribution of residuals is plotted. In table 6, some residuals statistics are presented, while in Fig.10, the histogram of the residuals is shown.

Figure 10: Histogram of regression's residuals

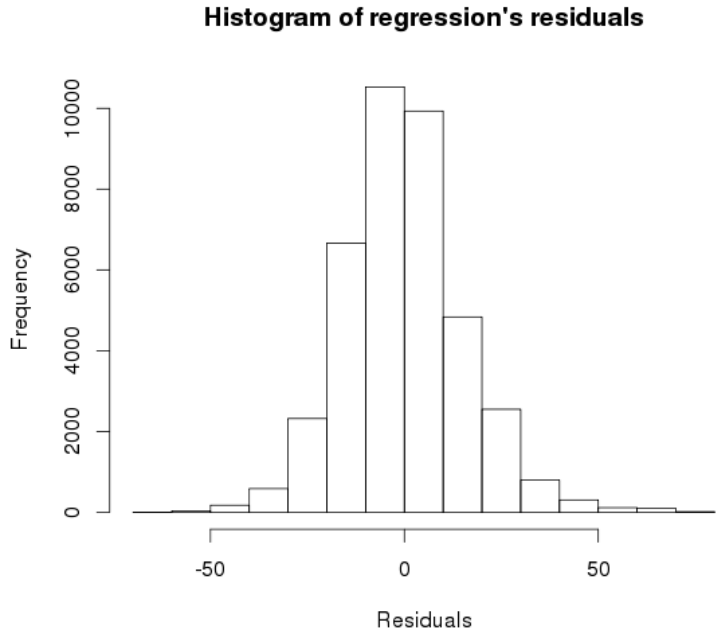
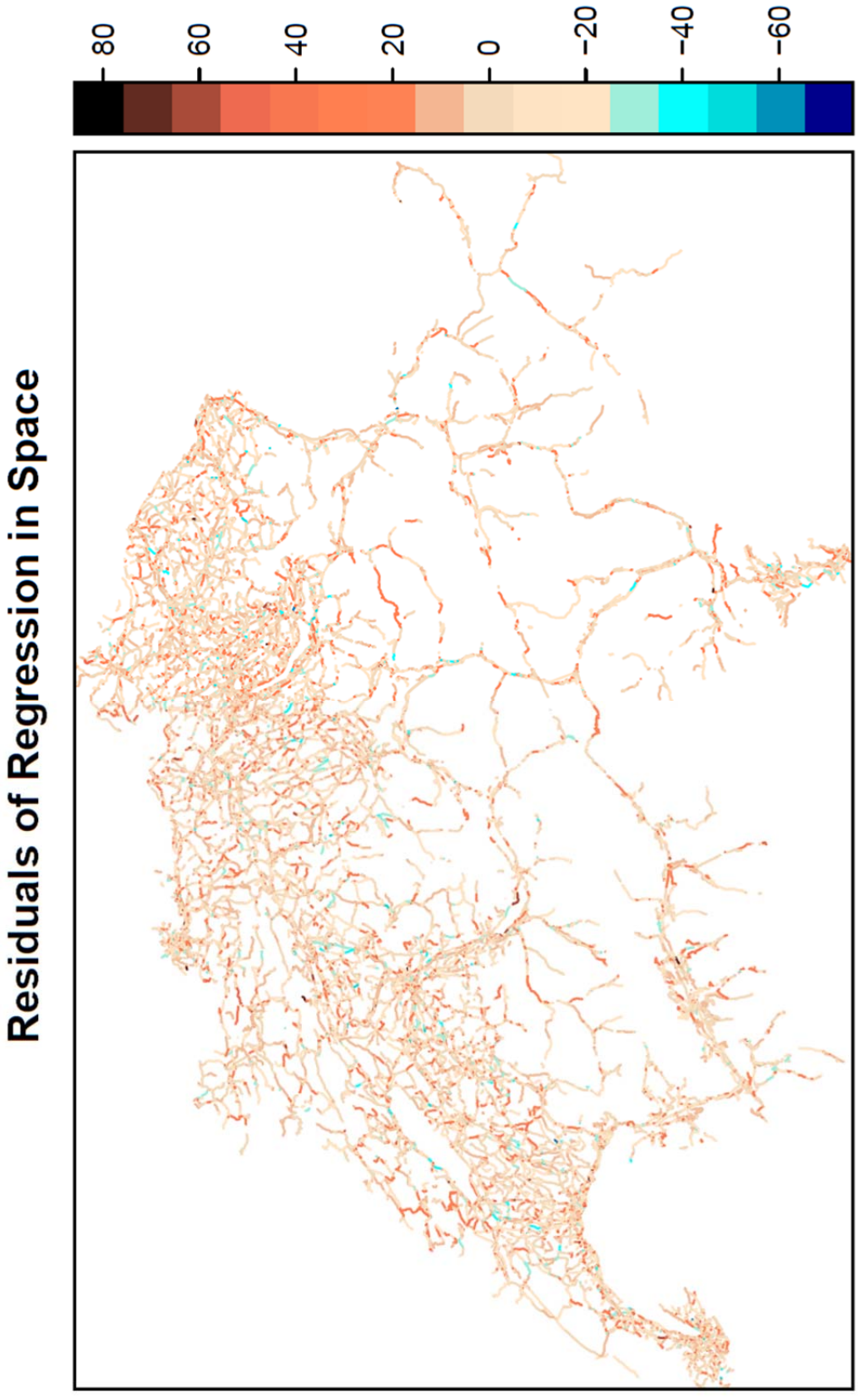


Table 6: Regression's residuals statistics

Min.	1Q	Median	3Q	Max	Std. Error
-65.441	-10.047	-0.668	8.599	76.004	15.55

As it can be seen in Fig.10, residuals seem to have a normal distribution. Positive residual values correspond to cases where the model underestimates speeds. The sign of the median is indicative of the existence of skewness with a direction to the left. Additionally, the larger magnitude of the first-quartile compared to the third quartile, indicates also a slight skew to the left in the data. Naturally, the maximum values of residuals raise some concerns however their magnitude can be attributed to wrongly matched links from the spatial join procedure, or can just be considered as outliers. A more detailed analysis is required to shed some light on the underlying cause of the existence of residuals and whether or not any pattern can be identified.

Figure 11: Spatial distribution of residuals



As it can be seen in the following graph (Fig.12), the cases of underestimation (positive residuals) are more often in the cases of links classified as primary, secondary, and tertiary. This finding makes apparent that the classification of the links of the OSM network might not be conducted in a fully consistent way, and the links within the same type might not be homogeneous with respect to their free flow speed, and subsequently their average daily speed. Consequently, the wrong classification might be the underlying cause of the sign of magnitude of the residuals. In Fig.13, no pattern can be identified regarding the existence of the residuals. In the cases of the rest types of links, the residuals range lies between -20 and 20 kilometers roughly.

Figure 12: Residuals of regression per road type

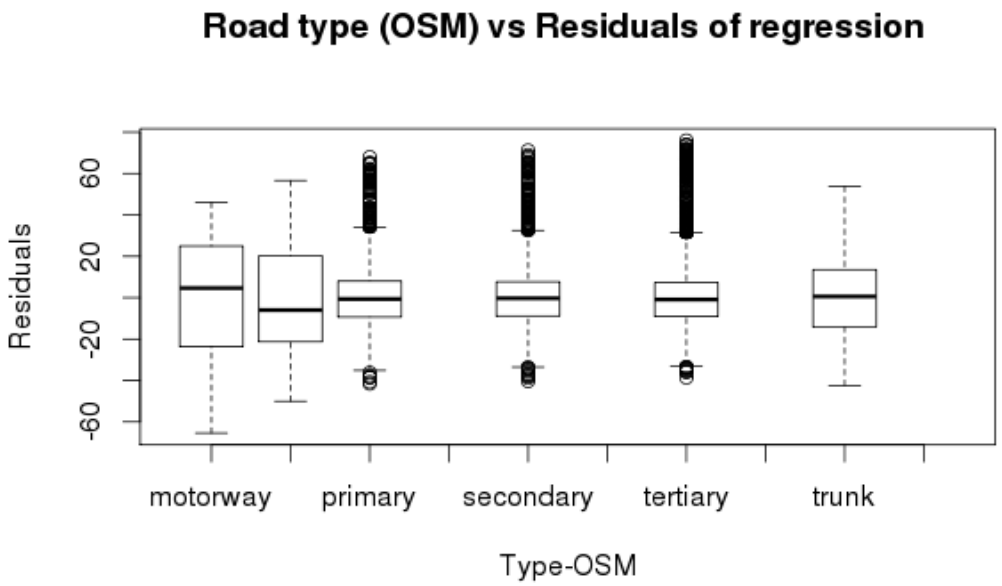
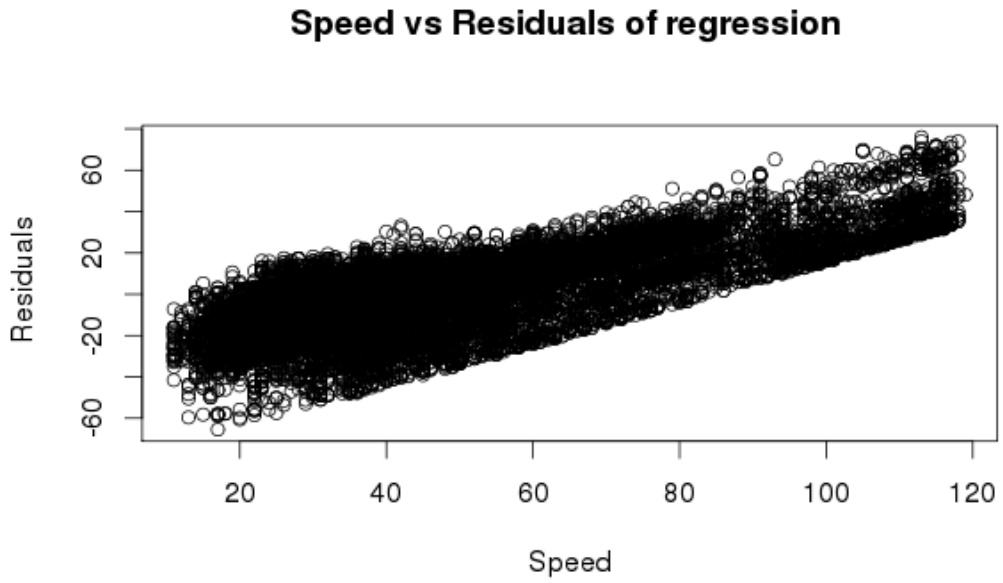


Figure 13: Residuals of regression against the actual speed



Residuals are also plotted against the speed per type in the figures that follow in order to check if any patterns can be identified in relationship to the classified type of the link. It can be seen in all of the plots that there is a positive correlation between the magnitude of the residual and the actual speed. In general, it appears that the model seems to fail to predict the speed of links that have significantly lower average daily speed than the corresponding average speed per type, either due to wrong classification of the links, or due to wrong spatial join of the links, or last due to congestion that is not captured sufficiently by the included variables in the model. The cases of overestimation for low speed values is more apparent in the cases of links with primary, secondary, and tertiary type. A finding which was also visible in Fig. 12.

Figure 14: Residuals of regression against the actual speed for motorway type

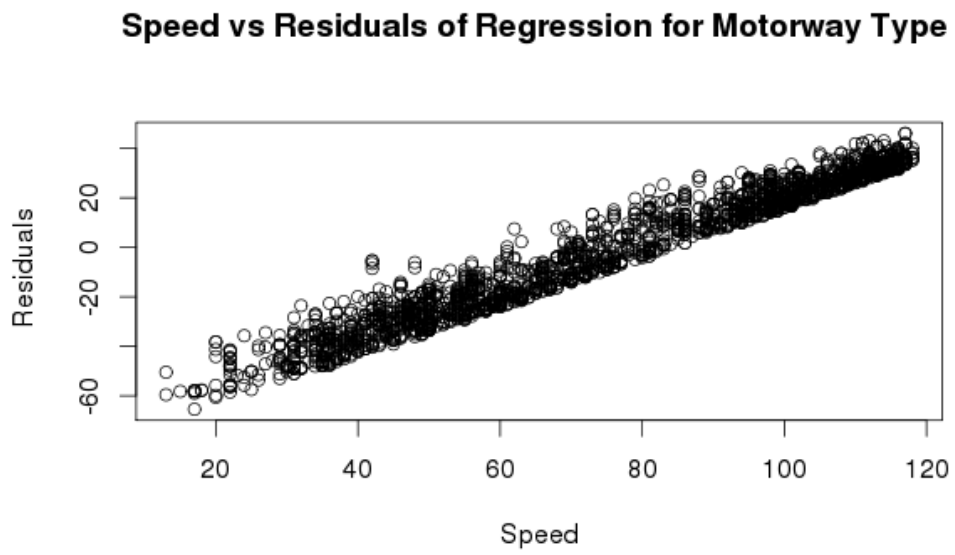


Figure 15: Residuals of regression against the actual speed for motorway links type

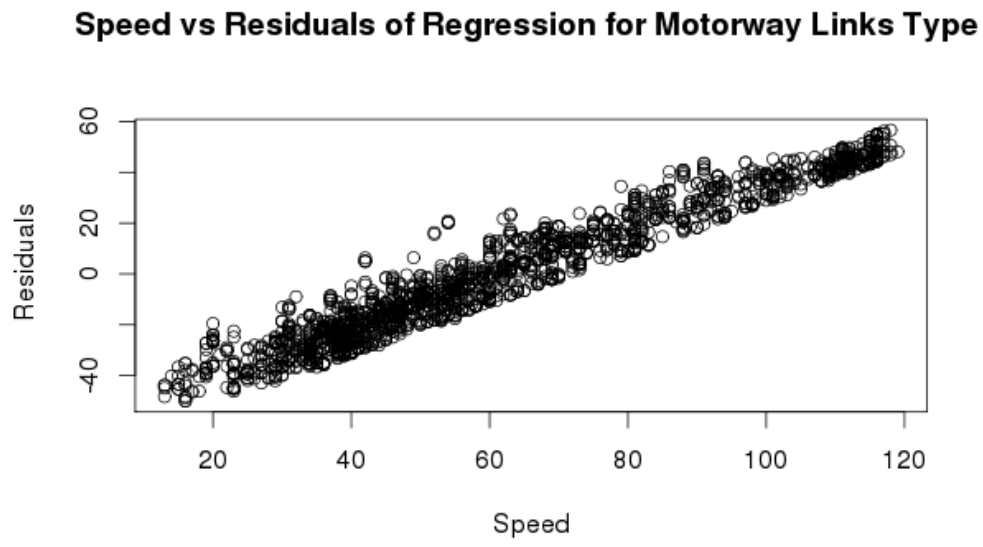


Figure 16: Residuals of regression against the actual speed for primary type

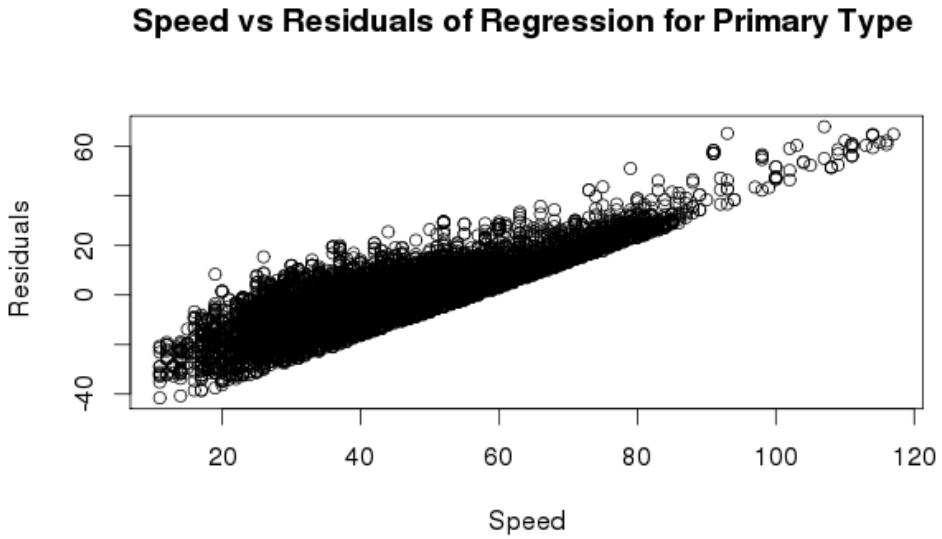


Figure 17: Residuals of regression against the actual speed for secondary type

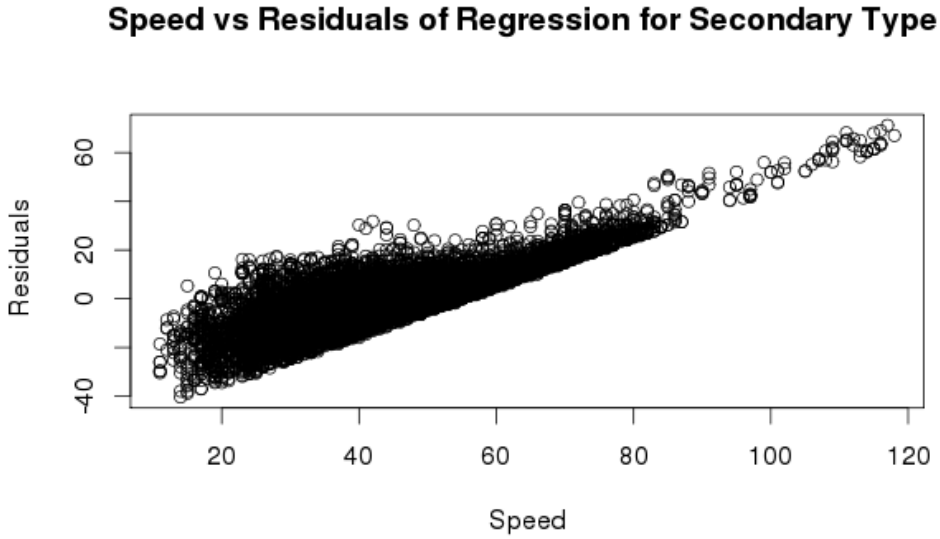


Figure 18: Residuals of regression against the actual speed for tertiary type

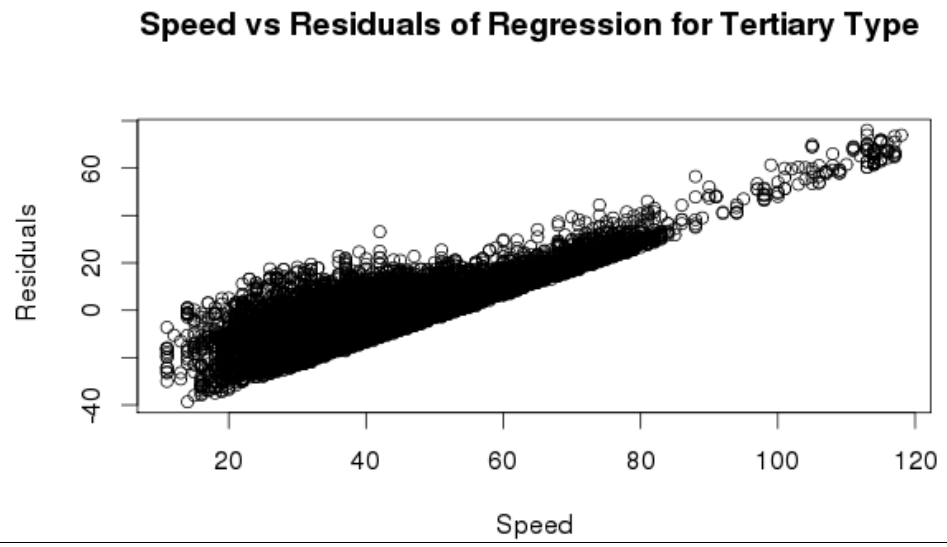
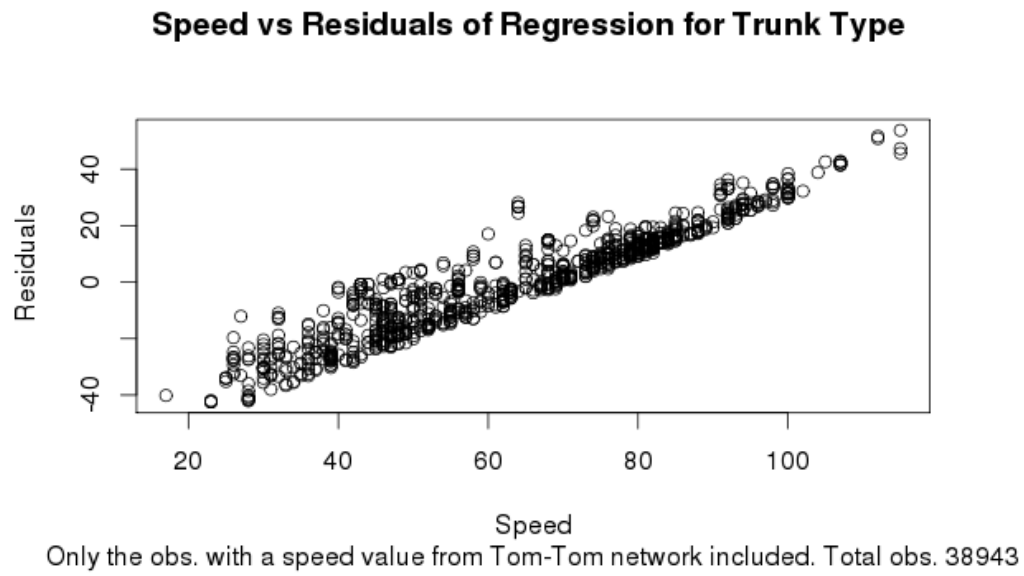


Figure 19: Residuals of regression against the actual speed for tertiary type



4 Application of the estimated model

The estimated model is applied to the full set of links included in the network of interest in order to predict their average daily speed for a typical weekday. As mentioned before, a sample of the links was used for the estimation of the model.

The results are delivered in a shapefile where the links' attributes named "predicted_speed" and "opp_pred_speed" respectively, correspond to the predicted speed values of interest, for both directions of each link. Apart from the predicted speed values, the variables used for the regression, are associated to the links, and included in the shapefile as well.

5 Conclusions and future work

In the previous sections, the estimation of a speed regression model that can be employed within an overall framework of a land-use and transport interaction tool, has been presented. The advantage of the coherent developed transport module is that it offers a direct method to predict the average speed values on the links of the network of interest, in a less cumbersome way than more advanced and complex transport demand models. The results presented that the method of regression modeling has the potential to be used for speed prediction purposes.

A set of improvements can be applied in the future to increase the predictive power of the estimated model. An obvious improvement constitutes the modification of the spatial join process to minimize the error caused by it. Ideally, speed measurements reported directly on the OSM network would constitute the most convenient solution. The quality of the data in the OSM network is another factor that affects in a significant way the results of the regression. The estimation of the model should also take place on another network, probably a commercial one, to assess the importance of missing values, and wrong registries (e.g. misclassification of links type). Future work can test the inclusion of more variables in the regression model that were unavailable at the time when the current study took place (e.g. number of lanes). Different formulations of the model that account for the spatial autocorrelation and the heterogeneity of the spatial data involved will consist an apparent improvement (spatial regression).

6 References

Hackney, J. K., M. Bernard, S. Bindra, and K. W. Axhausen (2007) Predicting road system speeds using spatial structure variables and network characteristics, *Journal of Geographical Systems*, 9 (4) 397–417.