A Universal Model of Commuting Networks

Maxime Lenormand, Sylvie Huet, Floriana Gargiulo, Guillaume Deffuant

A Universal Model of Commuting Networks

Maxime Lenormand[1], Sylvie Huet[1], Floriana Gargiulo[2] and Guillaume Deffuant[1]

[1] IRSTEA, LISC, 24 avenue des Landais, 63172 AUBIERE, France
(maxime.lenormand, sylvie.huet,guillaume.deffuant)@irstea.fr
[2] INED, 133 boulevard Davout, 75020 PARIS, France
floriana.gargiulo@gmail.com

Abstract. We test a recently proposed model of commuting networks on 80 case studies from different regions of the world (Europe and United-States) and with geographic units of different sizes (municipality, county, region). The model takes as input the number of commuters coming in and out of each geographic unit and generates the matrix of commuting flows betwen the geographic units. We show that the single parameter of the model, which rules the compromise between the influence of the distance and job opportunities, follows a universal law that depends only on the average surface of the geographic units. We verified that the law derived from a part of the case studies yields accurate results on other case studies. We also show that our model significantly outperforms the two other approaches proposing a universal commuting model (Balcan et al., 2009; Simini et al., 2012), particularly when the geographic units are small (e.g. municipalities).

1 Introduction

Commuting flows constitute the circulatory system of the modern societies: millions of people move every day from home to workplace and generate a network of socio-economic relationships wiring municipalities, counties or regions. These networks are the vector of several social and economic dynamics such as epidemic outbreaks, information flows, city development and traffic (Ortúzar and Willumsen, 2011; Balcan et al., 2009). Understanding their essential properties and reproducing them accurately is therefore a crucial issue for public health institutions, policy makers, urban development, infrastructure planners, etc. (De Montis et al., 2007, 2010)

In the abundant literature devoted to this challenge (see (Barthélemy, 2011; Rouwendal and Nijkamp, 2004) for reviews), the intuition a law inspired by gravitational attraction is widely accepted (Wilson, 1998; Choukroun, 1975): the number of commuters between two geographic units (cities, counties, regions...) is proportional to the product of the "masses" of each geographic unit (the population for example) and inversely proportional to a function of the distance between them. Unfortunately, numerous experiences showed that the

shape of the function of the distance and the basic parameter(s) of the model should be fixed in an ad-hoc manner for each case studies (de Vries et al., 2009; De Montis et al., 2007, 2010; Fotheringham, 1981). Therefore, it is impossible to generate commuting networks when data are lacking with this method.

In this paper, we show an universal law rules the single parameter of a recently proposed model (Gargiulo et al., 2012; Lenormand et al., 2012), which shows two main differences with the usual gravity law models:

- It takes as input the total number of commuters in and out from each geographic unit, instead of the population in usual gravity law models. It is hence more data demanding, but these data are widely available. From these data, the model reconstructs the whole network of flows between the geographic units.
- It builds the network progressively, considering dispatches commuters one by one in the different flows and it updates the virtual commuters in and out for each geographic unit after each virtual commuter out choice. This update allows to ensure the generated numbers of virtual commuters in and out for each unit are the same as the ones given by the observed data. The individual flow allocation follows a probability which increases with the number of commuters coming in the destination and decreases with the distance between the considered geographic units.

We test this model on 80 case-studies with geographic units of different sizes (for example in the same case-study the geographic unit can be either the municipality, the canton or the department, Fig. 1): Czech Republic (municipality scale, 1 region), France (municipality scale, 34 regions), France (canton scale, 14 regions + all France), France (département level (all France), Italy (municipality level, 10 regions), Italy (province level, 4 regions), USA (county level, 14 regions + all USA). We show that the single parameter of our model follows a simple universal law that depends only on the average area of the considered geographic units. This implies that, given the number of commuters in and out for each geographic unit and their average surface, we can derive the whole matrix of flows with a very good confidence.

Two other approaches (Balcan et al., 2009; Simini et al., 2012) can generate commuting networks only from population and job data. We show that our approach yields significantly more accurate results, especially when considering small geographic units (municipalities).

2 A simple model

The basic factors structuring the most commonly applied model of commuting networks, the "doubly-constrained" model (Wilson, 1998; Choukroun, 1975) include the number of commuters out and commuters in of the geographic units, and the distances between these units. The idea behind this choice is that individuals decide a work location taking into account the job offers and the distance

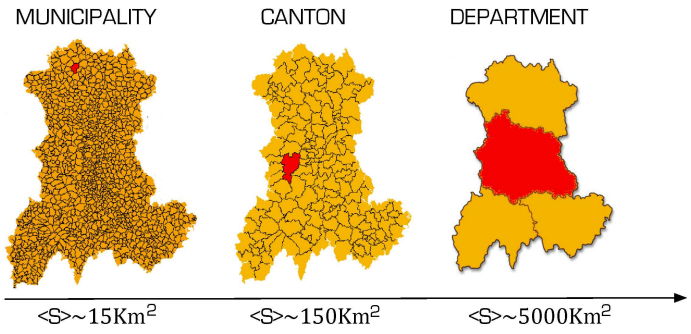

Fig. 1. Different sizes of unit for the French region Auvergne

to this work location. The distance is particularly important for everyday commuting, which is the most frequent case. We keep this basic setup, without adding any ingredient about the job market characteristics (professions, salary range, etc.). We propose a simple individual based procedure that allocates the individuals one by one to the different flows between geographical units, according to a probability that is inspired by gravitational models and that is updated after each allocation. More precisely, the probability for an individual, living in unit u_λ, to work in unit u_i is given by:

$$P_{\lambda \to i} = \frac{s_i^{in} e^{-\beta D_{\lambda i}}}{\sum_{k=1}^{N_{TOT}} s_k^{in} e^{-\beta D_{\lambda k}}} \tag{1}$$

where (s_i^{in}) is the number of commuters entering in unit u_i. $D_{\lambda i}$ is the Euclidian distance in meters between units u_λ and u_i (computable from the Lambert or GIS coordinates). This data is available in National Statistical offices (see appendix *Datasets* for more details), as well as s_λ^{out}, the number of commuters going out from unit u_λ. We choose a probability decreasing exponentially with the distance, in accordance with the investigations carried out in Lenormand et al. (2012) and with the literature on commuting network models. The impact of the distance is embedded in parameter β: For $\beta \to 0$ the probability is independent from the distance, while for high values of β, the probability tends to zero very rapidly when the distance increases, independently from the job offer.

We now describe the procedure in more details. The individuals live in a geographical area characterized by n territorial units, $u_\lambda \in \mathcal{U}$ with $\lambda \in |[1, n]|$, among which we want to generate the commuting network. Since a relevant part of our individuals can work outside the n units, especially those living close to the border of our area, to reduce the border effect (see (Lenormand et al., 2012)), we consider the job-search basin is an extended (EXT) area, composed by the

n residential units and m units surrounding the area (thus, we have $N_{TOT} = n + m$ units in total, $u_i \in \mathcal{U}^{EXT}$ with $i \in |[1, N_{TOT}]|$). The algorithm simulates individual searches for workplaces. At each time step we select unit u_λ at random among the residence units and one of its s_λ^{out} available commuters. We draw at random the working place u_i of this individual according to probabilities $P_{\lambda \to i}$. Then we decrement of one s_λ^{out} and s_i^{in}. Note that decrementing s^{in} and s^{out} at each step complicates significantly the derivation of an analytical expression of the model. The generated network is saved in matrix $\tilde{W} \in M_{n \times N_{TOT}}(\mathbb{N})$ where each entry $\tilde{W}_{\lambda i}$ represents the number of commuters between units $u_\lambda \in \mathcal{U}$ and $u_i \in \mathcal{U}^{EXT}$. The algorithm is summarized in Fig. 2.

Algorithm: Commuting generation model

Input : $D \in \mathrm{M}_{n \times N_{TOT}}(\mathbb{R})$, $s^{in} \in \mathbb{N}^{N_{TOT}}$, $s^{out} \in \mathbb{N}^n$, $\beta \in \mathbb{R}_+$
Output : $\tilde{W} \in \mathrm{M}_{n \times N_{TOT}}(\mathbb{N})$
  $\tilde{W}_{\lambda i} \leftarrow 0$
  while $\sum_{\lambda=1}^{n} s_\lambda^{out} > 0$ **do**
    Pick out at random $\lambda \sim A$ where
      $A = \{\mu | \mu \in |[1, n]|, s_\mu^{out} \neq 0\}$
    Pick out at random i from $|[1, N_{TOT}]|$
      with a probability $P_{\lambda \to i}$
    $\tilde{W}_{\lambda i} \leftarrow \tilde{W}_{\lambda i} + 1$
    $s_i^{in} \leftarrow s_i^{in} - 1$
    $s_\lambda^{out} \leftarrow s_\lambda^{out} - 1$
  end while
  return $\tilde{W}$

Fig. 2. Algorithm describing the network generation model

3 A universal law ruling the parameter β

We calibrated parameter β by minimizing the Kolmogorov-Smirnov (KS) distance between the observed and simulated distributions of commuting distances. We consider indeed that the distance distribution is an essential features that the model should reproduce. We checked this choice using the common part of commuters (CPC), based on the Sørensen index (Sørensen, 1948), which quantifies the similarity between the observed and simulated networks. Basically, the CPC computes which part of the commuting flows is correctly reproduced, on average, by the simulated network. The indicator varies between 0, when no agreement is found and 1 when the two networks are identical. We verified that the value of β that minimises the KS distance also maximises the CPC (see (Gargiulo et al., 2012; Lenormand et al., 2012) and in the appendix *Statistical*

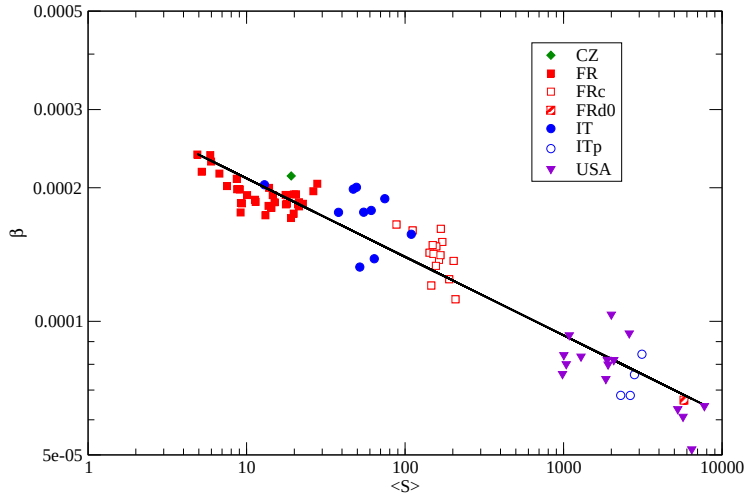

Fig. 3. Log-log scatter plot of the calibrated beta values in terms of average unit area (in km^2) for 80 regions; the line represents the regression line predicting the β value.

Tools for details). Moreover, we observed that the corresponding value of CPC is always higher than 0.70 with an average around 0.8 for all the case-studies, which shows a high similarity of the networks (see Fig. 4).

The next question is: How does the value of β vary with the global characteristics of the case-study? Actually, our results show that the optimal value of parameter β follows a universal rule depending only on the average surface of the geographic units. This rule is shown on Fig. 3, where the x-axis represents the average surface of the geographic units in the area ($\langle S \rangle$ in logarithm scale) and the y-axis the optimal β value (in logarithm scale). The linear regression in the log-log plane, shows a simple relation:

$$\beta = \alpha \langle S \rangle^{-\nu} \tag{2}$$

with $\alpha = 0.000315$, $\nu = 0.177$. The high value of the adjusted $R^2 = 0.92$ confirms the quality of the fit. We observe that β decreases with the average surface of the units $\langle S \rangle$, meaning that, when $\langle S \rangle$ is small (e.g. for French municipalities) the distance is more important in the commuting choice than when $\langle S \rangle$ is large (e.g. for regions or counties).

We use a cross-validation method to test the robustness of our estimation of the α and ν values and evaluate if it is possible to use them to generate commuting networks in new case studies. The dataset (including 80 case-studies) is randomly cut into two sets, called the training set (composed of 53 areas) and the testing set (composed of 27 areas). We use the training set to build a regression model giving the estimates of α and ν. From these estimates and

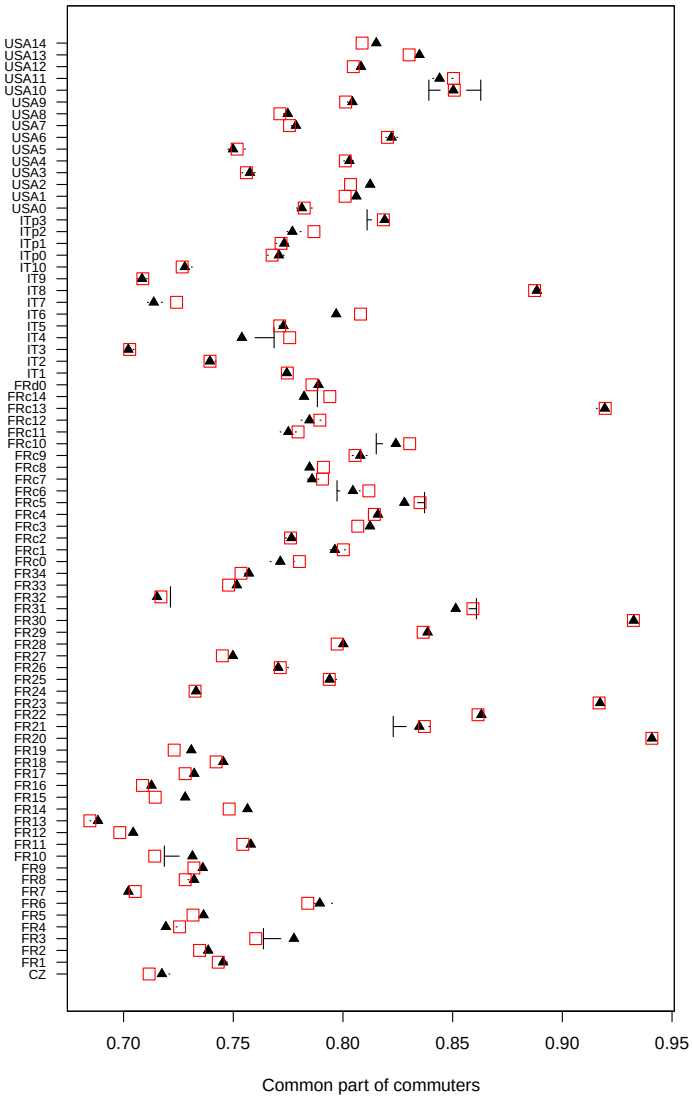

Fig. 4. Common part of commuters for the 80 regions. The squares represents the CPC obtained with the calibrated β value. The triangles represents the average CPC of the virtual networks build with the β values estimated by the cross validation procedure(over about 3500 estimated β values for each region) ; dark bars represent the minimum and the maximum CPC obtained from the network build with the estimated β but in most of the cases they are too closed to the average to be seen.

from equation 2, we compute β of the 27 regions of the testing set. We repeat the cross-validation process $10,000$ times obtaining about 3500 β estimations for each case study. Then we calculate the CPC for the value of β calibrated on the data (with the KS distance) and for each of the 3500 values obtained by the cross-validation method. Fig. 4 shows, for each case study, the CPC associated with the calibrated β, the average CPC obtained with the β values estimated from the cross-validation and the confidence interval defined by the minimum and the maximum values (but it is too small to be seen in most cases). The CPC obtained with the calibrated β value (black triangle) is almost the same as the average CPC obtained with the estimated β value (red square). We can observe that the average CPC obtained with the estimated β value is, for some areas, higher than the CPC obtained with the calibrated β value. It's possible that the common part of commuters are better with another β value because it's not the calibration criterion. Globally, we can conclude that the β values obtained with the log-linear model lead to the same values of the CPC indicator as the calibrated values. The method appears therefore fairly robust and can be used in other case studies with high confidence.

4 Discussion

We now discuss the interest of our proposal in comparison with two other important studies, Balcan et al. (2009) and Simini et al. (2012).

The objective of Balcan et al. (2009) is to generate a worldwide commuting network, and the model must deal with the wide variety of populations and surfaces of geographic units for which the data are available. To solve this difficulty, Balcan et al. (2009) project this data on ad-hoc units defined with a Voronoi diagram. They define their basic unit as a cell approximatively equivalent to a rectangle of 25 x 25 kilometers along the Equator. This allows them to calibrate their model because a unit is the same object whatever the country is. This is an interesting solution for generating a world-wide commuting network but it leads to an average commuting distance of 250 km which is much larger than the average distance of daily commuting (51 km in US, 28 km in UK and less in most of the other European countries). We expect our approach to take a better account of the heterogeneities in the greographic units.

The radiation model, proposed in Simini et al. (2012), is a universal approach for generating commuting networks: the commuting flow between two municipalities is a function of the cumulative job-opportunities at the distance between the two municipalities. The model has an elegant analytical solution and the average flow T^{ij} from unit i to unit j can be approximated by

$$\langle T_{ij} \rangle = \left(m_i \frac{N_c}{N} \right) \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{3}$$

where m_i and n_j are respectively the population of units u_i and u_j, N_c is the total number of commuters and N is the total population in the case-study region, and s_{ij} the total population in the circle of radius r^{ij} centred at u_i (excluding

the source and destination population). We implemented their analytical approximation and reproduced the graphs presented in their paper. Fig. 5 shows the comparison between the radiation model and ours in the US for inter-county commuting and in the French Auvergne region for inter-municipality commuting. We observe that in both cases our approach yields significantly better results. In particular the CPC measure for the radiation model for the inter-municipality commuting in Auvergne is 0.3, which indicates a poor matching with the data. To be fair, it should be reminded that our model uses more specific data (total number of commuters in and out of each geographic unit) than the radiation model, hence one could expect our results to be more accurate.

5 Conclusions

We propose a universal model of commuting network considering an individual choice for its place of work based on the principles of the gravity law, defining the attraction of a possible place of work as a function of its "approximated" or real job opportunities and of its distance from the place of residence. We generate the virtual commuting network for the residents of the units composing a case-study region. Following this individual decision function, a heuristic matching is done between each possible jobs of the various units (defined by the data on the commuters in for each unit) and each job seekers living in the unit (defined by the data on the commuters out of each unit). We show this model very relevant whatever the unit size is. It is in particular much more relevant than the other universal approaches since it allows building commuting between units of small size. This is more convenient to describe everyday commuting mostly corresponding to short distances.Moreover the stochastic property of our model allows to avoid considering small flows, especially those at short distance to a small unit, as deterministic. Once again, this last property is very relevant for virtual commuting among small units while at the same time informative on the confidence interval for large flows.

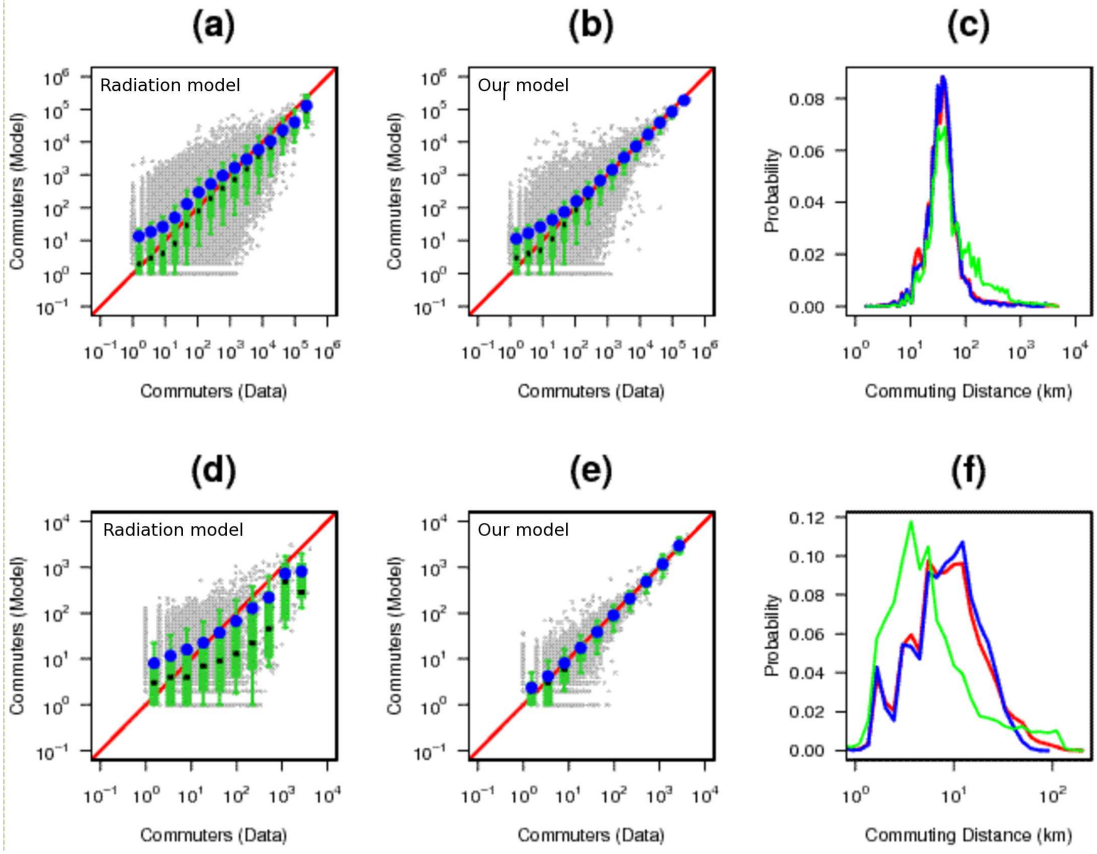

Fig. 5. Comparing the predictions of the radiation model with ours. Plots (a)-(c) US at county level, plots (d)-(f) Auvergne region (France) at municipality level). Plots (a),(b), (d), (e): comparison between the measured flows and the generated flows. Grey points are the scatter plot for each pair of counties. The black circles represent the average number of generated travelers in that bin. (a) and (d) plot the radiation model while (b) and (e) our model. Plots (c) and (f): commuting distance distributions of US (c) and Auvergne (f); the blue line represents the observed data, the red one the results of our model and the green one the results of the radiation model.

Bibliography

Balcan, D., Colizza, V., Goncalves, B., Hud, H., Ramasco, J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489.

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499:1–101.

Choukroun, J.-M. (1975). A general framework for the development of gravity-type trip distribution models. *Regional Science and Urban Economics*, 5(2):177–202.

De Montis, A., Barthelemy, M., Chessa, A., and Vespignani, A. (2007). The structure of interurban traffic: A weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924.

De Montis, A., Chessa, A., Campagna, M., Caschili, S., and Deplano, G. (2010). Modeling commuting systems through a complex network analysis: A study of the italian islands of sardinia and sicily. *The Journal of Transport and Land Use*, 2(3):39–55.

de Vries, J., Nijkamp, P., and Rietveld, P. (2009). Exponential or power distance-decay for commuting? an alternative specification. *Environment and Planning A*, 41(2):461–480.

Fotheringham, A. (1981). Spatial structure and distance-decay parameters. *Annals, Association of American Geographers*, 71(3):425–436.

Gargiulo, F., Lenormand, M., Huet, S., and Baqueiro Espinosa, O. (2012). Commuting network model: getting to the essentials. *Journal of Artificial Societies and Social Simulation*, 15(2):13.

Lenormand, M., Huet, S., and Gargiulo, F. (2012). Generating french virtual commuting network at municipality level.

Ortúzar, J. and Willumsen, L. (2011). *Modeling Transport*. John Wiley and Sons Ltd, New York.

Rouwendal, J. and Nijkamp, P. (2004). Living in two worlds: A review of home-to-work decisions. *Growth and Change*, 35(3):287–303.

Simini, F., Gonzalez, M. C., Maritan, A., and Barabasi, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, advance online publication:–.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34.

Wilson, A. G. (1998). Land-use/transport interaction models: Past and future. *Journal of Transport Economics and Policy*, 32(1):pp. 3–26.

6 Appendix: Datasets

Commuting data are usually provided by statistical offices in the form of origin-destination tables. We analyzed 80 regions from 7 differents datasets. In these description the outside are the units in $\mathcal{U}^{EXT}$ but not in $\mathcal{U}$.

6.1 Czech republic dataset at the municipality scale

This dataset is composed of the number of commuters between each couple of municipalities of South Moravia (A Czech Republic region)[3]. With this dataset we have built 1 region and its outside. The outside is composed of all the units of Czech Republic of except the ones belonging to the region. The region is identified with CZ.

6.2 French dataset at the municipality scale

This dataset is composed of the number of commuters between each couple of municipalities of France. The distance used is the Euclidean distance computed with the Lambert coordinates. With this dataset we have built 34 regions (French region or districts) and their outside composed by the neighboring French districts. This dataset is measured for the 1999 French Census by the French Statistical Institute, $INSEE$. They were kindly made available by the Maurice Halbwachs Center. These regions are identified from $FR1$ to $FR34$.

6.3 French dataset at the "canton" scale

This dataset is the same as the previous one at the "canton" scale (larger surface than municipality). The distance used is the Euclidean distance computed with the latitude and longitude. We used the longitude and latitude to build 14 regions and their outside. The outside is composed of all the units except that which compose the region. These region are identified from $FRc1$ to $FRc14$. We also used the complete French network (Noted $FRc0$) without outside.

6.4 French dataset at the "département" scale

This dataset is the same that the previous one at the "département" scale (larger surface than municipality and "canton"). The distance used is the Euclidean distance compute with the latitude and longitude. We built the complete French network (Noted $FRd0$) without outside.

6.5 Italy dataset at the municipality scale

This dataset is composed of the number of commuters between each couple of municipalities of Italy and the latitude and longitude of each municipality. We used the longitude and latitude to build 10 regions and their outside. The outside is composed of municipalities at a reasonable distance of the border of the region. These regions are identified from $IT1$ to $IT10$.

[3] Data are available online at http://www.czso.cz/xb/edicniplan.nsf/publ/13-6231-04-

6.6 Italy dataset at provincial level

The last dataset is the same that the previous one at the "provincia" scale (larger than municipality). We used the longitude and latitude to build 4 regions and their outside. The outside is composed of all the units except that which compose the region. These regions are identified from $ITp1$ to $ITp4$. We also used the complete Italian network (Noted $ITp0$) without outside.

6.7 United State of America dataset at the county scale

This dataset is composed of the number of commuters between each couple of counties of USA[4] and the latitude and longitude of each county[5]. We used the longitude and latitude to build 14 regions and their outside. The outside is composed of all the units except that which compose the region. These regions are identified from $USA1$ to $USA14$. We also used the complete USA network (Noted $USA0$) without outside.

[4] Available online at http://www.census.gov/population/www/cen2000/commuting/index.html
[5] Available online at http://www.census.gov/geo/www/gazetteer/places2k.html

Table 1. Description of the regions

Region	Number of units (region)	Number of units (outside)	Area (km^2)	Average Unit Area (km^2)	Standard deviation Unit area (km^2)	Number of commuters	Type of unit
CZ	43	630	35369	822.54	703.23	13309	Municipality
FR1	1310	3463	26013	19.86	12.49	295776	Municipality
FR2	1269	1447	27208	21.44	16.14	653710	Municipality
FR3	419	2809	5762	13.75	8.46	162370	Municipality
FR4	903	3081	8280	9.17	9.55	440961	Municipality
FR5	2296	2835	41309	17.99	21.30	700452	Municipality
FR6	261	3124	5175	19.83	10.46	69915	Municipality
FR7	185	1859	5167	27.93	18.71	12273	Municipality
FR8	1464	2467	25810	17.63	12.94	375363	Municipality
FR9	1842	4718	39151	21.25	14.76	624693	Municipality
FR10	3020	3845	45348	15.02	15.74	546162	Municipality
FR11	747	3169	16942	22.68	14.15	139481	Municipality
FR12	1786	3317	16202	9.07	7.46	268399	Municipality
FR13	1420	3536	12317	8.67	5.64	469335	Municipality
FR14	433	3914	6211	14.34	12.41	42690	Municipality
FR15	515	3808	5874	11.41	9.54	92053	Municipality
FR16	2339	3067	23547	10.07	7.51	547457	Municipality
FR17	260	1814	5565	21.40	13.15	23949	Municipality
FR18	1545	3046	27367	17.71	15.78	409116	Municipality
FR19	1948	1983	25606	13.14	12.94	375363	Municipality
FR20	36	1245	176	4.89	3.28	973173	Municipality
FR21	262	1543	2284	8.72	6.62	618741	Municipality
FR22	185	1707	1246	6.74	3.83	526600	Municipality
FR23	47	1234	245	5.21	3.03	642092	Municipality
FR24	377	2283	3525	9.35	7.44	183504	Municipality
FR25	195	2338	3718	19.07	17.66	41600	Municipality
FR26	547	449	4116	7.52	15.87	65469	Municipality
FR27	163	353	4299	26.37	27.53	163445	Municipality
FR28	327	2788	4781	14.62	9.76	178828	Municipality
FR29	102	2031	609	5.97	4.21	45185	Municipality
FR30	40	783	236	5.90	4.28	655200	Municipality
FR31	196	1597	1804	9.20	6.04	518321	Municipality
FR32	463	2588	5229	11.29	8.03	59963	Municipality
FR33	433	2728	6004	13.87	9.07	75561	Municipality
FR34	286	2088	5857	20.48	13.36	49815	Municipality
FRc0	3146	0	540241	171.72	99.90	12193161	Canton
FRc1	1062	2084	173797	163.65	91.23	2576191	Canton
FRc2	523	2623	58366	111.60	114.44	4141190	Canton
FRc3	226	2920	33041	146.20	70.56	661813	Canton
FRc4	160	2986	25044	156.52	75.47	379668	Canton
FRc5	55	3091	7847	142.67	71.64	100783	Canton
FRc6	869	2277	131174	150.95	96.62	2876966	Canton
FRc7	2088	1058	351073	168.14	94.18	5020735	Canton
FRc8	100	3046	20246	202.46	161.41	346184	Canton
FRc9	600	2546	113905	189.84	103.57	1392498	Canton
FRc10	302	2844	26627	88.17	77.64	2485733	Canton
FRc11	906	2240	142619	157.42	100.21	2457521	Canton
FRc12	1500	1646	250676	167.12	99.00	3526558	Canton
FRc13	32	3114	6653	207.91	145.33	63538	Canton
FRc14	506	2640	75603	149.41	85.63	1537545	Canton
FRd0	94	0	540250	5747.35	1957.11	3548178	Département
IT1	377	0	24090	63.90	61.89	225351	Municipality
IT2	395	201	24157	61.16	77.51	415530	Municipality
IT3	1002	2020	54918	54.81	71.37	1282522	Municipality
IT4	201	507	14964	74.45	82.42	332176	Municipality
IT5	204	1005	10567	51.80	55.68	297749	Municipality
IT6	51	506	5582	109.45	101.52	72270	Municipality
IT7	2000	4001	98693	49.35	60.97	3005328	Municipality
IT8	186	1023	2412	12.97	15.25	408777	Municipality
IT9	1510	4004	71167	47.13	58.08	1757794	Municipality
IT10	705	3008	26809	38.03	41.62	455568	Municipality
ITp0	99	0	277220	2800.20	1619.86	1567576	Province
ITp1	50	49	131773	2635.45	1401.23	945194	Province
ITp2	30	69	93666	3122.21	1599.56	325279	Province
ITp3	20	79	45854	2292.72	1128.38	538752	Province
USA0	3108	0	8070785	2596.78	3437.29	34104128	County
USA1	1015	2093	1876151	1848.42	916.86	6554650	County
USA2	103	3005	101411	984.57	341.47	707091	County
USA3	54	3054	306284	5671.93	4488.99	736084	County
USA4	2011	1097	4169235	2073.21	1786.40	15287520	County
USA5	202	2906	404093	2000.46	1994.32	9423862	County
USA6	504	2604	949238	1883.41	1041.57	2473662	County
USA7	806	2302	4234740	5254.02	5626.18	5438917	County
USA8	352	2756	2723212	7736.40	7741.02	4305008	County
USA9	1507	1601	2877429	1909.38	1517.28	10919198	County
USA10	13	3095	14123	1086.37	343.73	123923	County
USA11	32	3076	205989	6437.17	4105.95	69129	County
USA12	1004	2104	1292835	1287.68	563.79	10458777	County
USA13	207	2901	207785	1003.79	352.24	1819403	County
USA14	301	2807	312955	1039.72	394.71	2365275	County

7 Appendix: Statistical tools

7.1 Calibration

In each case study area we build a normalized histogram $P(d)$ describing the probability that a commuter travels a certain distance d. This histogram shows a typical log-normal shape in all the studied areas, with a peak varying from case to case. Each β value produces a different distance histogram: low values of β generate uniform distance distributions, while high values give exponentially decreasing structures.
To obtain the β value for each case study, we minimize the Kolmogorov-Smirnov distance between the histogram for the observed and the simulated data:

$$D_{KS}(\beta) = \sup_d |P^{OBS}(d) - P^{SIM}(d, \beta)| \tag{4}$$

As we can observe in Fig. 6, the Kolmogorov-Smirnov distance presents a clear minimum in correspondence of an optimal β value that is different in different areas.

7.2 Validation

After we found the optimal value of the parameter we must verify the efficiency of the model in reproducing the data. To evaluate the validation procedure we use two origin-destination matrices (Table 2), the observed one $Y \in M_{(n+1)\times(n+1)}(\mathbb{N})$ and the simulated one $\tilde{Y} \in M_{(n+1)\times(n+1)}(\mathbb{N})$. Y can be easily obtained by difference with the total number of in-commuters $(s_\lambda^{in})_{1\leq\lambda\leq N_{TOT}}$, the total number of out-commuters $(s_i^{out})_{1\leq i\leq n}$ and the light grey table of the Table 3 correponding to W. To compare Y and $\tilde{Y}$ we use as statistical indicator the Sørensen similarity index, an indicator usually used to evaluate the similarity of content of different samples for ecological problems. In our specific case we specifically call the index "Common part of commuters" and we define it in the following way:

$$CPC(Y, \tilde{Y}) = \frac{2NCC(Y, \tilde{Y})}{NC(Y) + NC(\tilde{Y})} \tag{5}$$

where $NCC(Y, \tilde{Y})$ is the number of commun commuters between the two sets:

$$NCC(Y, \tilde{Y}) = \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \min(Y_{ij}, \tilde{Y}_{ij}) \tag{6}$$

$NC(Y)$ and $NC(\tilde{Y})$ are respectively the number of commuters in the observed and simulated sets:

$$NC(Y) = \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} Y_{ij} \qquad NC(\tilde{Y}) = \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \tilde{Y}_{ij} \tag{7}$$

This indicator varies between 0 if the simulation values never reproduce the observed ones to 1 if the perfect agreement is realized.
In the lower plot of Fig. 6 we can observe that the model reaches the best performance in reproducing the original results (higher Søørensen index) exactly in th

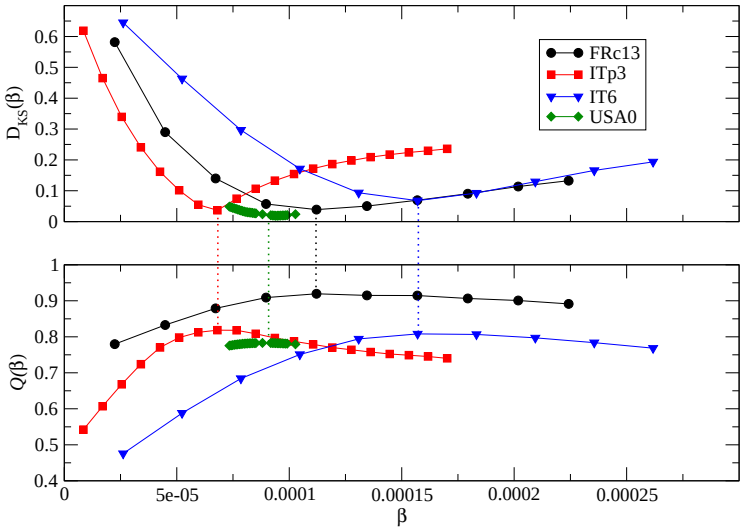

Fig. 6. Upper plot: Kolmogorov-Smirnov distance between the real and the simulated distance distributions as a function of beta, for some case study areas. Lower plot: Sørensen index as a function of β. Each point is the result of 100 replicas of the generation process

Table 2. Origin-destination table; The light grey table represents the commuters living and working in the region for each municipality of the region; The grey columns represent the out-commuters living in the region and working outside (Out.) for each municipality of the region; The grey line represents the in-commuters working in the region and living outside (Out.) for each municipality of the region; The dark grey line(column) represents the total number of out(in)-commuters for each municipality of the region.

WP / **RP**	u_1	...	u_j	...	u_n	**Out.**	**Total**
u_1	0	...	Y_{1j}	...	Y_{1n}	$Y_{1\text{out}}$	s_1^{in}
...	...	...	...	...	...	...	...
u_i	Y_{i1}	...	Y_{ij}	...	Y_{in}	$Y_{i\text{out}}$	s_i^{out}
...	...	...	...	...	...	...	...
u_n	Y_{n1}	...	Y_{nj}	...	0	$Y_{n\text{out}}$	s_n^{out}
Out.	Y_{out_1}	...	Y_{out_j}	...	Y_{out_n}		
Total	s_1^{in}	...	s_j^{in}	...	s_n^{in}		

Table 3. Origin-destination table from the region to the region and the outside; The light grey table represents the commuters living (place of residence RP) and working (place of work WP) in the region for each municipality of the region; The grey table represents the commuters living (place of residence RP) in the region and working (place of work WP) outside of the region.

WP / **RP**	u_1	...	u_i	...	u_n	u_{n+1}	...	u_N^{TOT}
u_1	0	...	W_{1i}	...	W_{1n}	W_{1n+1}	...	$W_{1N^{TOT}}$
...	...	...	...	...	...	...	...	...
u_λ	$W_{\lambda 1}$	...	$W_{\lambda i}$	...	$W_{\lambda n}$	$W_{\lambda n+1}$	...	$W_{\lambda N^{TOT}}$
...	...	...	...	...	...	...	...	...
u_n	W_{n1}	...	W_{ni}	...	0	W_{nn+1}	...	$W_{nN^{TOT}}$