Author-produced version of the paper Lenormand M, Huet S, Gargiulo F, Deffuant G (2012)
A Universal Model of Commuting Networks. PLoS ONE 7(10): e45985. doi:10.1371/journal.pone.0045985
Available at http://www.plosone.org

A Universal Model of Commuting Networks

Maxime Lenormand[1,*], Sylvie Huet[2], Floriana Gargiulo[3], Guillaume Deffuant[4]
1 LISC, Irstea, Clermont-Ferrand, France
2 LISC, Irstea, Clermont-Ferrand, France
3 CAMS, CNRS, Paris, France
4 LISC, Irstea, Clermont-Ferrand, France
∗ **E-mail: maxime.lenormand@irstea.fr**

1 Abstract

We show that a recently proposed model generates accurate commuting networks on 80 case studies from different regions of the world (Europe and United-States) at different scales (e.g. municipalities, counties, regions). The model takes as input the number of commuters coming in and out of each geographic unit and generates the matrix of commuting flows between the units. The single parameter of the model follows a universal law that depends only on the scale of the geographic units. We show that our model significantly outperforms two other approaches proposing a universal commuting model [1,2], particularly when the geographic units are small (e.g. municipalities).

2 Introduction

Billions of people move everyday from home to workplace and generate networks of socio-economic relationships that are the vector of social and economic dynamics such as epidemic outbreaks, information flows, city development and traffic [1, 3]. Understanding the essential properties of these networks and reproducing them accurately is therefore a crucial issue for public health institutions, policy makers, urban development, infrastructure planners, etc. [4, 5]. This challenge is the subject of an intensive scientific activity (see [6, 7] for reviews), in which the analogy of the gravitational attraction inspires a majority of approaches [8, 9]: the number of commuters between two geographic units (cities, counties, regions...) is supposed 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 experiments showed that the optimum function and parameter values vary a lot with the case studies [4, 5, 10, 11]. This situation is not satisfactory because when one wants to generate a particular commuting network without having the total origin destination matrix of commuting, no practical heuristic is available for choosing the adequate type of function and parameter values. This paper addresses this problem.

We consider a recently proposed model [12,13], differentiating itself from the usual gravity law models in two main features:

- It takes as input the total number of commuters in and out from each geographic unit. With this starting point, the model focuses directly on the influence of the distance between geographic units on the commuting probability. The model is data demanding, but these data are widely available.
- It builds the network progressively, allocating commuters one by one in the different flows, according to probabilities that increase with the number of commuters coming in the destination and decrease with the distance between the origin and destination. These probabilities are updated after each allocation.

Our model is close to the traditional doubly-constrained gravity model [8, 9], but it is more flexible and less data demanding. Indeed, the doubly constrained model and the methods used to solve it require a closed network of commuters: they cannot take into account commuting links outside the considered

geographical units. Our individual based stochastic approach overcomes this problem and can deal with the usually available data of total number of commuters in and out of geographic units.

We test this model on 80 case-studies with geographic units of different scales. For example in the same case-study the geographic unit can be either the municipality, the canton or the department, (see an example on Figure 1). More precisely, the case studies include: Czech Republic (municipality scale, 1 case-study), France (municipality scale, 34 case-studies), France (canton scale, 15 case-studies including whole France), France (département scale one case-study (whole France), Italy (municipality scale, 10 case-studies), Italy (province scale, 4 case-studies), USA (county level, 15 case-studies including whole USA). For a detailed description of the datasets see the Supplementary Information *Datasets*.

We show that the single parameter of our model follows a simple universal law that depends only on the average surface of the considered geographic units. This implies that, given the number of commuters in and out of each geographic unit and their average surface, we can derive the whole matrix of flows with a very good confidence.

Two other approaches [1,2] claim to catch universal properties of commuting networks. We show that our model yields significantly more accurate results, especially for case-studies with small geographic units (e.g. municipalities).

3 The model

We consider the basic double-constrained model setup, without adding any ingredient about the job market characteristics (professions, salary range, etc.). Instead of solving analytically the optimisation problem, we use an individual based procedure that allocates virtual individuals one by one in the different flows between geographic units, according to a probability that is updated after each allocation.

This individual based approach can deal with less constrained data than the doubly-constrained gravity model that requires the total number of commuters in to be equal to the total number of commuters out. In other words the doubly contrained model can only deal with the flows between the considered geographic units; it cannot take into account the commuting links with destinations outside the case study area. This is a problem when only the numbers of commuters in and out the geographic units are available (and not the complete matrix of the commuting flows), because the data do not distinguish between the flows inside and outside the case study area. It is therefore difficult to estimate the correct data to take as input to the doubly-constrained model in this case. Our approach is more flexible and overcomes this difficulty. It does not require that the total number of commuters in and out to be equal (for more details see [13]), hence it can easily use directly the usually available data on the number of commuters in and out of each geographic unit.

Let s_i^{out} and s_j^{in} be respectively the global number of commuters starting from unit u_i and the global number of commuters arriving in unit u_j. These numbers are initialised from data and then they are progressively modified by the procedure. More precisely, at each step we select unit u_i such that $s_i^{out} > 0$ at random, and we consider a virtual commuter starting from u_i. We draw at random the working place u_{j^*} of this individual among all possible destinations u_j according to probabilities $P_{i\to j}$:

$$P_{i\to j} = \frac{s_j^{in} e^{-\beta D_{ij}}}{\sum_{k=1}^{N} s_k^{in} e^{-\beta D_{ik}}} \tag{1}$$

where D_{ij} is the Euclidian distance in meter between units u_i and u_j (computable from the Lambert or GIS coordinates). Having drawn u_{j^*}, we decrement of one s_i^{out} and $s_{j^*}^{in}$. Note that decrementing s^{in} and s^{out} at each step complicates significantly the derivation of an analytical expression of the model. We chose a probability decreasing exponentially with the distance, in accordance with the investigations carried out in [13] and with the literature on commuting network models. The importance of the distance in the commuting choices is embedded in parameter β: for $\beta \to 0$ the probability tends to be independent

from the distance, while for high values of β, the probability tends to zero very rapidly when the distance increases, independently from the number of commuters arriving in the units.

To reduce the border effect (see [13]), we consider the job-search basin in an extended (EXT) area, composed by the n residential units and m units surrounding the area. Thus, we have n units which are commuting origins and $N = n + m$ units that are commuting destinations. The generated network is saved in matrix $\tilde{T} \in M_{n\times N}(\mathbb{N})$ where each entry $\tilde{T}_{ij}$ represents the number of commuters between units u_i and u_j. The algorithm is summarized in Figure 2.

4 A universal law ruling parameter β

The model depends on a single parameter ruling the importance of the distance in commuting choice. We show that this parameter can be derived as a function of the scale of the problem, independently from the socio-geographical location of the case study area. This opens the possibility to reconstruct the commuting flows (origin-destination matrix) when they are not provided.

We calibrated parameter β by maximising the common part of commuters (CPC), based on the Sørensen index [14].

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

with:

$$NCC(T,\tilde{T}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\min(T_{ij},\tilde{T}_{ij}) \qquad NC(T) = \sum_{i=1}^{n}\sum_{j=1}^{n}T_{ij} \tag{3}$$

where T is the observed origin-destination matrix and $\tilde{T}$ is the simulated one. This is a similarity measure based on the Sørensen index in ecology computing which part of the commuting flows is correctly reproduced, on average, by the simulated network. It varies between 0, when no agreement is found, and 1, when the two networks are identical. We priviledged this indicator because of its direct interpretation. Indeed, when $NC(T) \simeq NC(\tilde{T})$ (it is the case for our model), the CPC represents the percentage of commuting connection correctly located (i.e. with the right pair origin - destination). Moreover, we tested on all case studies that the results obtained with the MAE, the RMSE or CPC[1] are equivalent (see the Supplementary Information *Other indicators* for more details). As an example on the $FR1$ case study, Figure 3 shows that the same β value maximizes the CPC and minimizes the MAE. In this Figure we can also note that the CPC is very sensitive to β and that its value does not vary much with the different replicas of the stochastic solving process.

Moreover, in order to have an idea of the improvement of the model compared with complete randomness, we have computed the CPC of a random model where the probabilities presented in Eq. (1) are uniform ($P_{i\to j} = \frac{1}{n}$, where n is the number of units). As shown on the Figure 4 we obtained an average CPC around 0.1. For our model, the CPC is always higher than 0.7 with an average around 0.8, which can be interpreted as 70 to 80 % of correctly predicted commuting connections.

Our goal is to derive the value of β from some easily available global characteristics of the case-study, giving the possibility to reconstruct the commuting flows when they are not available. Figure 5 gives strong evidence of such a universal relation.

The x-axis represents the average surface of the geographic units of the case-study ($\langle S\rangle$ 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:

$$\beta = \alpha\langle S\rangle^{-\nu} \tag{4}$$

[1]We have also shown in [12, 13] that the value of β yielding the maximum CPC also yields the maximum similarity between observed and simulated commuting distance distributions

with $\alpha = 3.15 \cdot 10^{-4}$ and $\nu = 0.177$. α corresponds to the β value for the unitary surface 1 km^2. The high value of the adjusted $R^2 = 0.92$ confirms the quality of the linear model. 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 municipalities in France) the distance is more important in the commuting choice than when $\langle S \rangle$ is large (e.g. for regions or counties).

We now evaluate the robustness of our estimation of α and ν using a common statistical procedure: the cross-validation. The cross-validation aims at evaluating the potential error of using the β value derived from the regression model intsead of deriving this value by optimisation for a new case study. This procedure repeats a large number of times the following steps: define a sub-sample of the total sample of case studies, derive a regression model of β from this sub-sample, for each case study that do not belong to the sub-sample, derive β from this regression model and compare the corresponding CPC with the value of β directly calibrated on the complete origin - destination data. The dataset (including 80 case-studies) is randomly cut into two sets, called the training set (comprising 53 case-studies) and the test set (composed of 27 case-studies). We build a regression model on the training set, providing α and ν, from which we derive estimates of β for each of the 27 case-studies of the testing set. We have 27 estimations of β using the relation 4 where α and ν are obtained from the random sub-sample of 53 case-studies. We repeat this process $10,000$ times obtaining 270,000 estimations of β (uniformly distributed over the 80 case-studies) corresponding to about $\frac{270,000}{80} = 3,375$ estimations of β for each case study. Then we calculate the average, minimum and maximum CPC for each of these values of β, and we compare them with the CPC obtained with value of β directly calibrated on the data.

Figure 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 in most cases (red square). Globally, we can conclude that the β estimated with the log-linear model and the calibrated β lead to very similar CPCs and also very similar MAE and the RMSE as shown in the appendix *Other indicator*. The method appears therefore fairly robust and this gives confidence for using it with the value of β derived from our loglog regression in new cases studies.

5 Comparaison with other universal derivations of commuting networks

Two other different approaches, [1] and [2], claim also to provide a universal derivation of commuting networks. The objective of [1] 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, the authors project these data on ad-hoc units defined with a Voronoi diagram. They define their basic unit as a cell approximately 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. 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. For example for the USA case study the average distance of daily commuting is about 68 km for the observed network and about 64 km for the simulated network obtained with our algorithm. For the Auvergne (France) case study at municipality scale the average distance of daily commuting is about 12 km for the observed network and about 11 km for the simulated one.

In the radiation model, proposed in [2], the commuting flow between two geographic units is a function of the cumulated population in a circle at the distance between the two units. The model has an elegant

analytical solution and the average flow T_{ij} from unit u_i to unit u_j can be approximated by

$$\langle T_{ij} \rangle = \left(m_i \frac{P_c}{P} \right) \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{5}$$

where m_i and n_j are respectively the population of units u_i and u_j, P_c is the total number of commuters and P 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. Figure 6 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. Moreover, as shown on Figure 4, the average CPC for the radiation model on all the case studies is around 0.4, and lower for all case studies than the one obtained with our approach.

However, 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. Therefore, to be fair with the radiation model we implemented a modified version of this model using the number of out and in commuters of each units. This new approximation is presented in equation 6 where s_{ij} the total number of in-commuters in the circle of radius r_{ij} centred at u_i (excluding the source and destination).

$$\langle T_{ij} \rangle = s_i^{out} \frac{s_i^{out} s_j^{in}}{(s_i^{out} + s_{ij})\left(s_i^{out} + s_j^{in} + s_{ij}\right)} \tag{6}$$

As shown on Figure 4, this new model reaches an average CPC around 0.5 which is higher than the original radiation model but still significantly lower than the results obtained with our model. Using the MAE and the RMSE leads to the same conclusions (see the appendix *Other indicator* for more details).

6 Discussion

The power law of our model's single parameter β with the average area of the case study geographic units, is surprising to us because of the high variety in our case studies in terms of scale, number of units, number of commuters and surface areas. For instance the Auvergne region in France is rural with a population density of about 50 hab./km^2 whereas the New York City region is very urban with a population density of about 6500 hab./km^2. As far as we know, this is the first time that a single model is shown to fit such diverse group of datasets.

We show that our approach outperforms the radiation model and that the difference of input data plays a minor role in this superiority. This superiority is not due to our particular treatment of the border effects either. Indeed, we could check our approach outperforms the radiation model also on particular case studies (e.g. on islands such as Corsica) where this border effect does not play. We can conclude that the accuracy of our model comes from a proper use of the number of commuters in and out of each geographic unit and an adequate choice of the function of the distance.

The results of the cross validation procedure give a good confidence in the robustness of this law. However, we have to admit that, despite their diversity, our 80 case studies come all from western industrialised countries. Therefore it will be important to check the validity of our law on case studies coming from other continents and less industrialised countries. Moreover, we use a very rough approximation of the distance between the geographic units with the Euclidian distance between the unit centroids. More accurate approximations of this distance would certainly improve the results. Finally, we also intend to apply our approach to commuting networks inside urban areas because many cities of the world show an impressive growth and an increasing part of commuting takes place within them [15]. An important issue in our perspective is to check if our law holds at this scale.

Acknowledgments

References

1. Balcan D, Colizza V, Goncalves B, Hud H, Ramasco J, et al. (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: 21484-21489.

2. Simini F, Gonzalez MC, Maritan A, Barabasi AL (2012) A universal model for mobility and migration patterns. Nature advance online publication: –.

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

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

5. De Montis A, Chessa A, Campagna M, Caschili S, 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: 39-55.

6. Barthélemy M (2011) Spatial networks. Physics Reports 499: 1-101.

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

8. Wilson AG (1998) Land-use/transport interaction models: Past and future. Journal of Transport Economics and Policy 32: pp. 3-26.

9. Choukroun JM (1975) A general framework for the development of gravity-type trip distribution models. Regional Science and Urban Economics 5: 177-202.

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

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

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

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

14. 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.

15. Roth C, Kang SM, Batty M, Barthélémy M (2011) Structure of urban movements: Polycentric activity and entangled hierarchical flows. PLoS ONE 6.

7 Figure Legends

Author-produced version of the paper Lenormand M, Huet S, Gargiulo F, Deffuant G (2012)
A Universal Model of Commuting Networks. PLoS ONE 7(10): e45985. doi:10.1371/journal.pone.0045985
Available at http://www.plosone.org

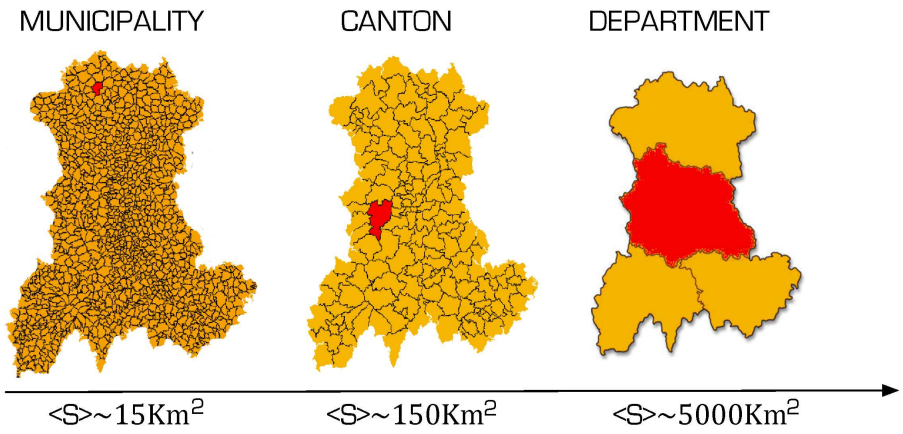

Figure 1. Three scales of geographic units (Auvergne region, France)

Algorithm: Commuting generation model

Input : $D \in \mathrm{M}_{n\times N}(\mathbb{R})$, $s^{in} \in \mathbb{N}^N$, $s^{out} \in \mathbb{N}^n$, $\beta \in \mathbb{R}_+$
Output : $\tilde{T} \in \mathrm{M}_{n\times N}(\mathbb{N})$
$\tilde{T}_{ij} \leftarrow 0$
while $\sum_{k=1}^{n} s_k^{out} > 0$ **do**
Pick at random $i \in |[1,n]|$, such that $s_i^{out} \neq 0$
Pick at random j from $|[1,N]|$
with a probability $P_{i\to j}$
$\tilde{T}_{ij} \leftarrow \tilde{T}_{ij} + 1$
$s_j^{in} \leftarrow s_j^{in} - 1$
$s_i^{out} \leftarrow s_i^{out} - 1$
end while
return $\tilde{T}$

Figure 2. Algorithm describing the network generation model

Author-produced version of the paper Lenormand M, Huet S, Gargiulo F, Deffuant G (2012)
A Universal Model of Commuting Networks. PLoS ONE 7(10): e45985. doi:10.1371/journal.pone.0045985
Available at http://www.plosone.org

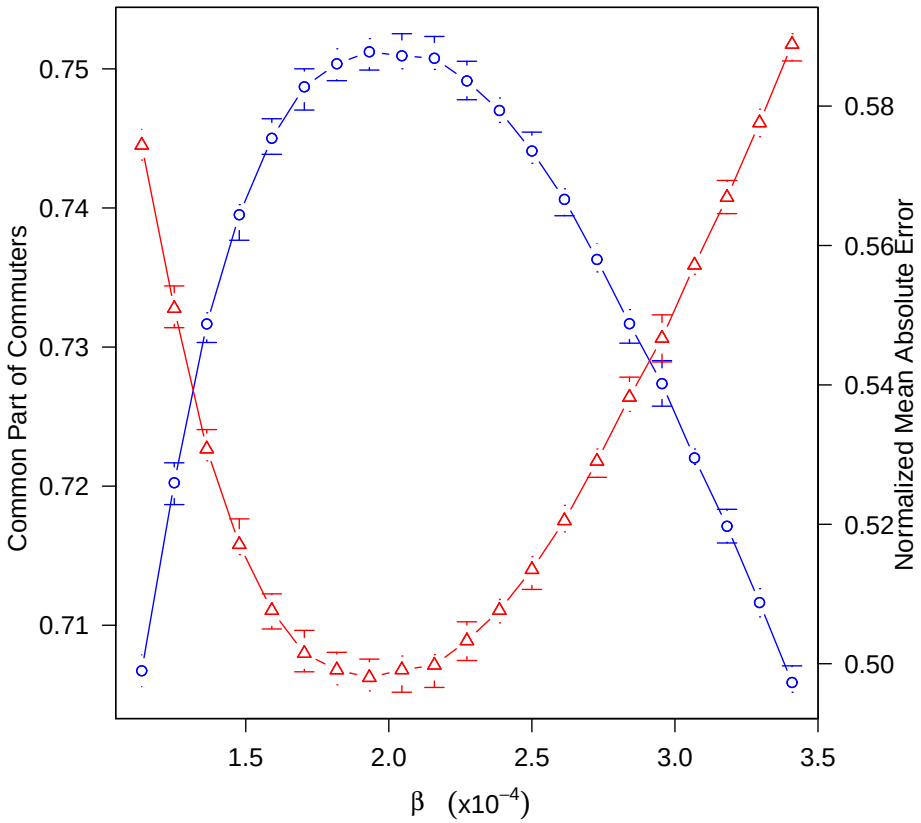

Figure 3. Plot of the average CPC (blue circle) and the average NMAE (red triangle) in term of β for 10 replications of the model for the Auvergne case study (FR1). The error bars represent the minimum value and maximum value obtain over the 10 replications.

Author-produced version of the paper Lenormand M, Huet S, Gargiulo F, Deffuant G (2012)
A Universal Model of Commuting Networks. PLoS ONE 7(10): e45985. doi:10.1371/journal.pone.0045985
Available at http://www.plosone.org

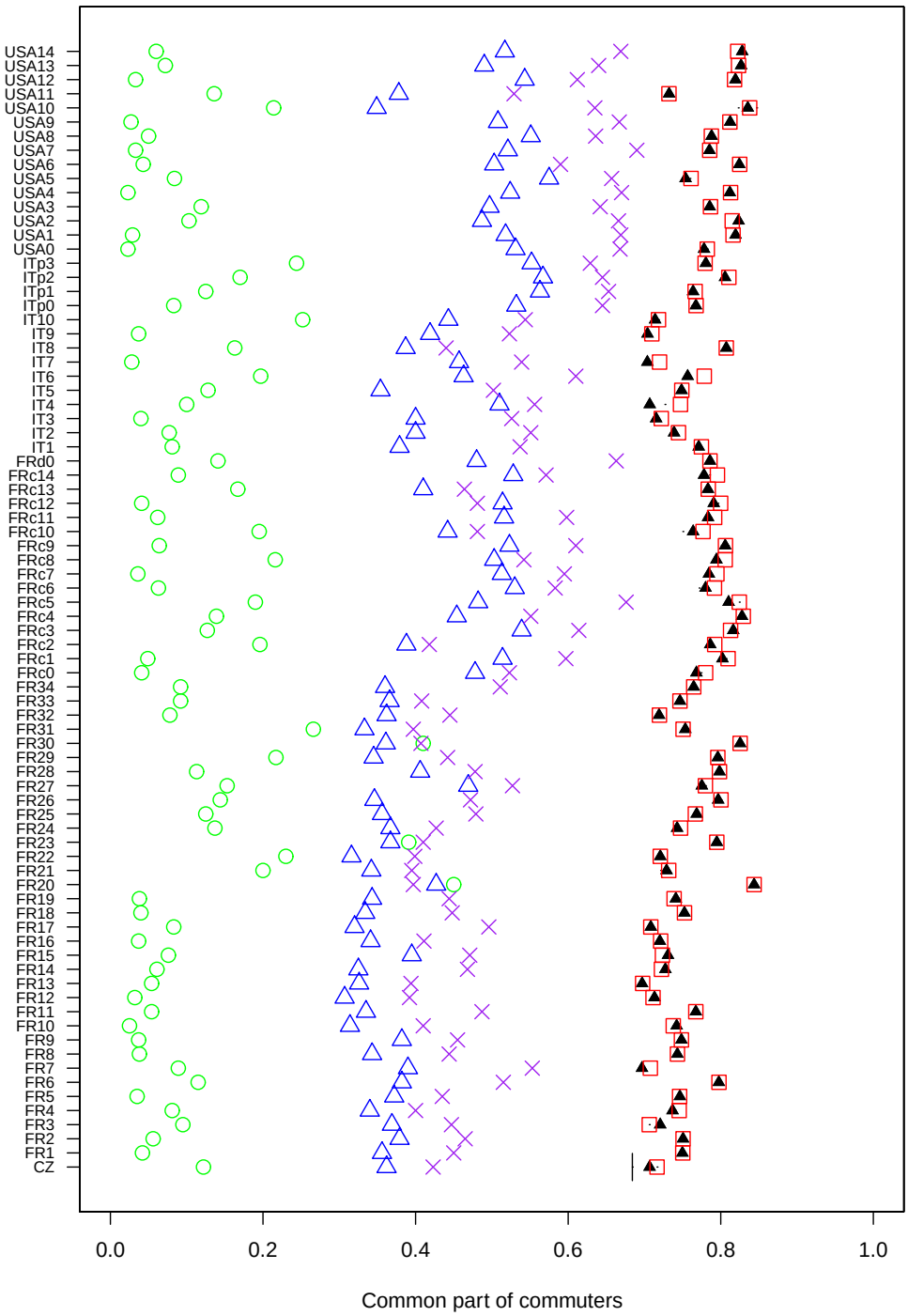

Figure 4. Common part of commuters (CPC) for the 80 case-studies. The red squares represent the CPC obtained with the value of β optimised from data on the case-study network. Black plain triangles represent the average CPC obtained with β values estimated with the rule linking β and the average surface of the units obtain with the cross-validation; Dark bars represent the minimum and the maximum CPC obtained with the estimated β but in most cases they are too close to the average to be seen. The green circles represent the CPC obtained with the random model. The blue triangles represent the CPC obtained with the radiation model. The purple crosses represent the CPC obtained with the modified version of the radiation model.

Author-produced version of the paper Lenormand M, Huet S, Gargiulo F, Deffuant G (2012)
A Universal Model of Commuting Networks. PLoS ONE 7(10): e45985. doi:10.1371/journal.pone.0045985
Available at http://www.plosone.org

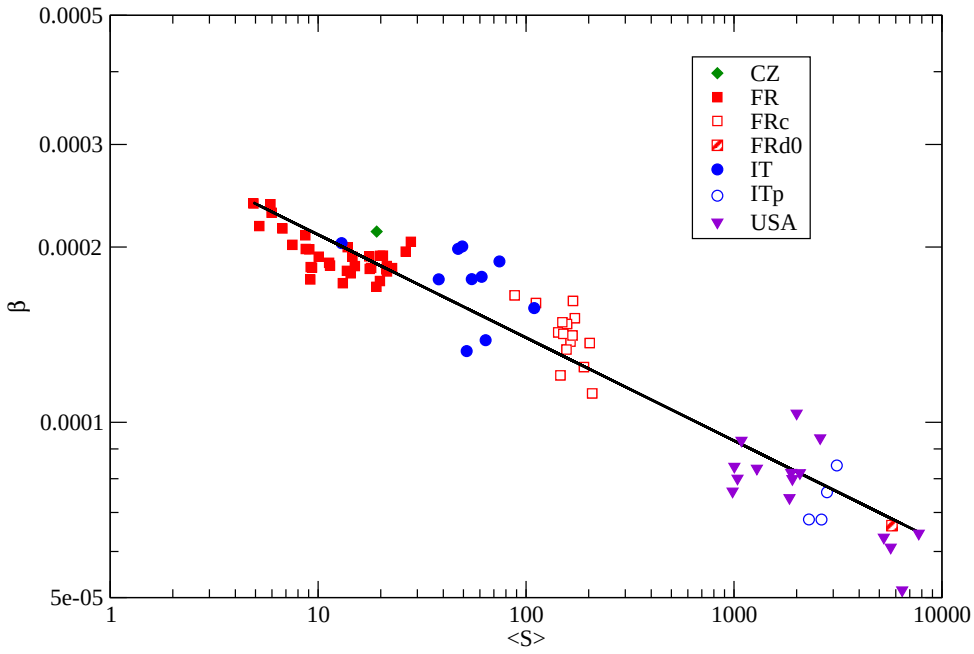

Figure 5. Log-log scatter plot of the calibrated β values in terms of average surface of the geographic units for 80 case-studies; the line represents the regression line predicting β. The surface is made non–dimensional by the unitary surface 1 km^2.

Author-produced version of the paper Lenormand M, Huet S, Gargiulo F, Deffuant G (2012)
A Universal Model of Commuting Networks. PLoS ONE 7(10): e45985. doi:10.1371/journal.pone.0045985
Available at http://www.plosone.org

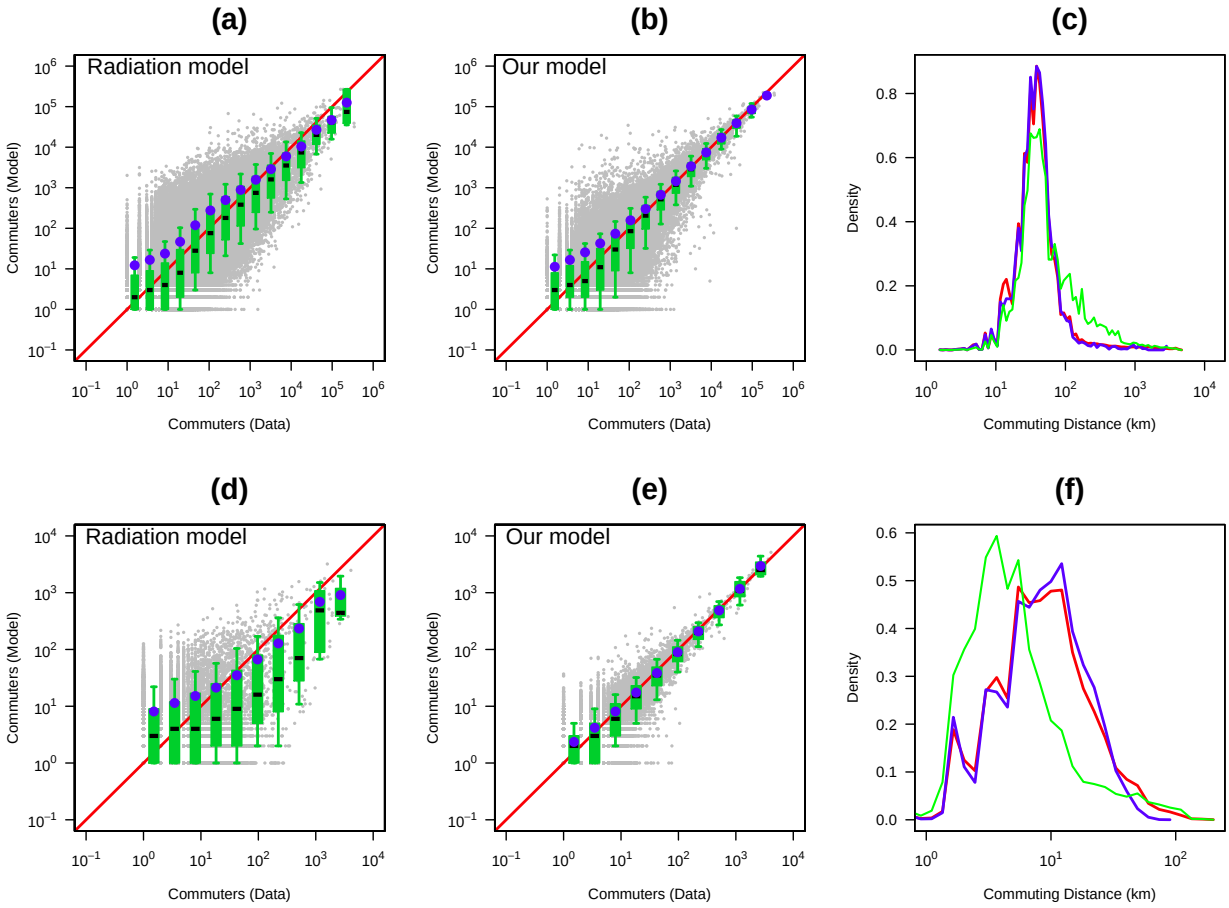

Figure 6. Comparing the predictions of the radiation model with ours for two case studies, the first row ((a)-(c)) for $USA0$ (USA at county scale) and the second row ((d)-(f)) for $FR1$ (Auvergne region, France at municipality scale). Plots (a), (b) , (d) and (e): Comparison between the observed (Census) and the simulated (model) non-zero flows. Grey points are the scatter plot for each pair of units. The boxplots (D1, Q1, Q2, Q3 and D9) represent the distribution of the number of simulated travelers in different bins of number of observed travelers. The blue circles represent the average number of simulated travelers in the different bins. Plots (c) and (f): Commuting distance distributions (km) (i.e. Probability for a commuters of the region to commut at a distance d). 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.