Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?[☆]

A. Briant[*,a], P.-P. Combes[b,1], M. Lafourcade[a,c]

[a] *Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris, France*
[b] *University of Aix-Marseille, GREQAM, 2 rue de la Charité, 13236 Marseille cedex 02, France*
[c] *University of Paris-Sud 11, ADIS, 54 Boulevard Desgranges, 92331 Sceaux Cedex, France*

Abstract

This paper evaluates, in the context of economic geography estimates, the magnitude of the distortions arising from the choice of a specific zoning system, which is also known as the Modifiable Areal Unit Problem (MAUP). We undertake three standard economic geography exercises (the analysis of spatial concentration, agglomeration economies, and trade determinants), using various French zoning systems differentiated according to the size and shape of their spatial units. While size might matter, especially when the dependent variable of a regression is not aggregated in the same way as the explanatory variables and/or the zoning system involves large spatial units, shape does so much less. In any case, both dimensions are of secondary importance compared to specification issues.

Key words: MAUP, concentration, agglomeration, wage equations, gravity.
JEL: R12, R23, C10, C43, O18.

1. Introduction

Most empirical work in economic geography relies on scattered geo-coded data that are aggregated into discrete spatial units, such as cities or regions. However, the aggregation of spatial dots into boxes of different size and shape is not benign regarding statistical inference. The sensitivity of statistical results to the choice of a particular zoning system is known as the Modifiable Areal Unit Problem (hereafter MAUP). Surprisingly, economists paid little attention to this problem up until recently.[2] Our main objective here is to

[☆] This paper was prepared as part of the PREDIT research contract 04-MT-5036, whose financial support is gratefully acknowledged. The authors are also involved in the Marie Curie Research Training Network 'TOM - Transnationality of migrants' funded by the European Commission (contract No: MRTN-CT-2006-035873). We thank Alain Sauvant and Micheline Travet (French Ministry of Transport) for providing us with the trade data, and Sébastien Roux for his very kind cooperation on wage data. We are indebted to Yasushi Asami, Andrew Clark, Gilles Duranton, Laurent Gobillon, Keith Head, Thierry Mayer, Daniel McMillen, Henry Overman, Matthew Turner, as well as two anonymous referees and William Strange for insightful comments. Conference and seminar participants in Kyoto, Toronto, Barcelona, Lyon and Pau also provided us with very useful feedback.

[*] Corresponding author.
Email addresses: anthony.briant@m4x.org (A. Briant), ppcombes@univmed.fr (P.-P. Combes), lafourcade@pse.ens.fr (M. Lafourcade)
URL: http://www.vcharite.univ-mrs.fr/pp/combes/ (P.-P. Combes), http://www.pse.ens.fr/lafourcade/ (M. Lafourcade)
[1] Also affiliated with the CEPR and Paris School of Economics. Financial support from CNRS is gratefully acknowledged.

[2] Two noticeable exceptions are Holmes and Lee (forthcoming) and Menon (2008).

assess whether differences in results across empirical studies are really sparked by economic phenomena in the process under scrutiny, or rather just by different zoning systems. We first investigate whether changes in either the *size* (equivalently the number) of spatial units, or their *shape* (equivalently the drawing of their boundaries) alter any of the estimates that are usually computed in the economic geography literature. Second, we address the important question of whether distortions due to the MAUP are large compared to those resulting from specification changes.

Disentangling these two effects is essential for policy. For instance, much work has tried to check empirically whether agglomeration enhances economic performance at the scale of countries, European regions, U.S. states or even smaller spatial units such as U.S. counties or French employment areas. The magnitude of the estimates differs between papers, but we do not know whether this reflects zoning systems or real differences in the extent of knowledge spillovers, intermediate input linkages, and labor-pooling effects on firm productivity. The resulting economic policy prescriptions regarding cluster-formation strategies will be affected accordingly. In the same vein, a large body of literature has evaluated the degree of spatial concentration, but does not check whether the conclusion that some industries are more concentrated than others results from the chosen zoning system or from more fundamental differences in the size of agglomeration and dispersion forces across industries at different spatial scales.

This paper is based on three standard empirical questions in economic geography, although many others could have been considered.[3] We start by evaluating the degree of spatial concentration under three types of French zoning systems (administrative, grid and partly random spatial units) and by comparing the differences between concentration measures (Gini vs. Ellison and Glaeser) with those between zoning systems. We then turn to regression analysis as not only is the measure of any spatial phenomenon likely to be sensitive to the MAUP, but also its correlation with other variables. We estimate the impact of employment density on labor productivity and compare the magnitude of agglomeration economies across zoning systems and econometric specifications. Finally, we run gravity regressions. We study how changes in the size and shape of spatial units affect the elasticities of trade flows within France with respect to both distance- and information-related trade costs.

All of these empirical exercises suggest that, when spatial units remain small, changing their size only slightly alters economic geography estimates, and changing their shape matters even less. Both distortions

[3]For comparison purposes, we use the same specifications as those typically found in the literature (see Combes, Mayer, and Thisse, 2008b), even though we do not necessarily think that they are the most apt.

are secondary compared to specification issues. More caution should be warranted with zoning systems involving large units, however. The MAUP is obviously less pervasive when data variability is preserved from one scale to another. When moving from dots to boxes, specific attention should be devoted to the following key points: 1- the size of boxes in comparison with the original dots, 2- the way data are aggregated, i.e. averaging or summation, 3- the degree of spatial autocorrelation in the data. The MAUP is less jeopardizing when data are spatially-autocorrelated and averaged, as is the case in wage regressions. By way of contrast, the MAUP is more challenging when variables in a regression are not computed under the same aggregation process. In gravity regressions for instance, moving from one scale to another requires a summation of trade flows on the left-hand side, whereas distance is averaged on the right-hand side.

The remainder of the paper is organized as follows. Section 2 provides a simple illustration of the possible size- and shape-dependency of spatial statistical inference, along with a data simulation exercise. Section 3 lists the zoning systems for which our estimations are carried out. As a first sensitivity test, section 4 is dedicated to the study of French spatial concentration patterns. Sections 5 and 6 investigate the extent to which changing econometric specifications and zoning systems affect the size and significance of wage and trade determinants respectively. Section 7 concludes and suggests further lines of research.

2. The Modifiable Areal Unit Problem : A Quick Tour

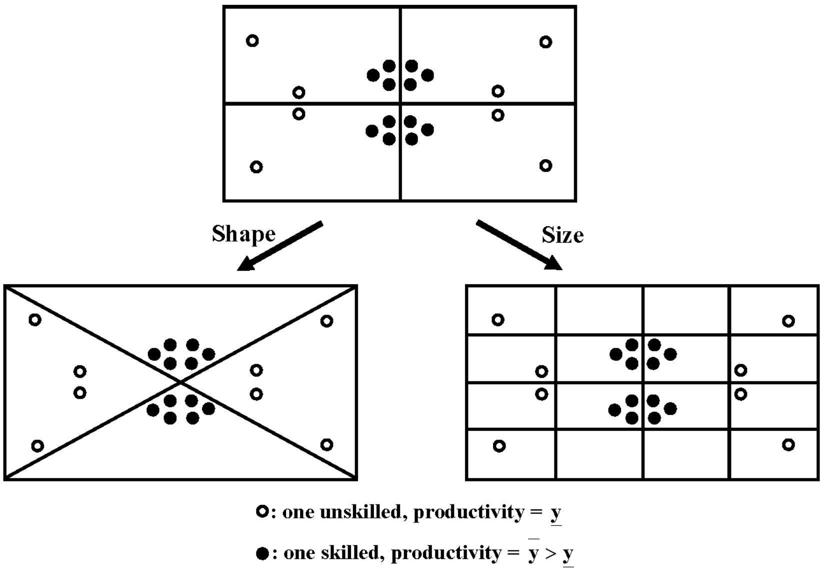
The Modifiable Areal Unit Problem is a longstanding issue for geographers. In their seminal contribution, Gehlke and Biehl (1934) were the first to emphasize that simple statistics such as correlation coefficients could vary tremendously across zoning systems. They note that, in the United States, the correlation between male juvenile delinquency and the median equivalent monthly housing rent increases monotonically with the size of spatial units. Openshaw and Taylor (1979) pursued this line of investigation and, drawing on correlations between the percentage of Republican voters and the percentage of the population over sixty, standardize what they called the "Modifiable Areal Unit Problem".[4]

2.1. A simple illustration of the MAUP

Spatial statistics may vary along two dimensions: firstly, the level of aggregation, or the *size* of spatial units, and secondly, at a given spatial resolution, the drawing of their boundaries, or their *shape*. Figure 1 illustrates these two related issues via the employment density - labor productivity relationship.

[4] See Fotheringham and Wong (1991) for an extended review of the earliest MAUP contributions.

Figure 1 – The size and shape issues



Black points display the location of skilled workers, whose individual productivity is denoted $\overline{y}$, while empty dots stand for unskilled workers, with productivity $\underline{y} < \overline{y}$. In the top figure, space is divided into four rectangles, each consisting of three skilled and two unskilled workers. The spatial distribution of workers across units is uniform and average productivity is the same across units. To illustrate the shape effect, consider the bottom-left figure. Spatial concentration emerges here, with two clusters of six high-skilled workers and two clusters of four low-skilled workers. Average productivity is higher in the former due to the spatial sorting of labor skills. Hence, agglomeration economies, defined here as the positive correlation between productivity and employment density, are zero in the first zoning system but positive in the second. We now turn to the size effect. In the bottom-right figure, we consider smaller rectangles with the same proportions as in the top figure. Spatial concentration is also found here, but the relationship between productivity and density is less marked than in the bottom-left case. Indeed, the difference in productivity between low- and high-productivity regions remains the same (except for empty boxes), whereas the density gap is higher in the bottom-right case. Hence, the extent and scope of agglomeration economies change with the size and shape of units, even though the underlying spatial information -the location and productivity of workers- remains the same.

The question we pursue in this paper is hence twofold. How much does moving from a particular zoning system to another alter the perception of an economic phenomenon? And how does this alteration vary accordingly to whether information is summed or averaged under this aggregation process? Section 2.2 provides a first clue to these questions, drawn from a simple simulation exercise.

2.2. *Mean and variance distortions: a first illustration with simulated data*

A number of authors have provided detailed analyses of the MAUP based on simulated data. According to Arbia (1989), both size and shape distortions are minimized (although never eliminated) under two restrictive conditions that are rarely met in practice: the exact equivalence of sub-areas (in terms of size, shape and neighboring structure) and the absence of spatial autocorrelation. In a subsequent work, Amrhein (1995) carries out a simulation exercise where he draws $10,000$ values from a randomly-generated variable and allots *randomly* each of these values to a Cartesian address within a unit square. In doing so, the value at one address is independent of the values at contiguous addresses and there is no spatial autocorrelation.[5] The author then divides the unit-square into respectively 100, 49 and 9 equally-sized sub-squares. Finally, he aggregates the information by averaging the values assigned to each sub-square. In line with Arbia (1989), he concludes that, under the strong assumption of random allocation, means do not display any pronounced size and shape effects and the changes in variances are only driven by the fall in the number of units.[6] Based on Canadian Census data, Amrhein and Reynolds (1997) further show that the distortions of simple statistics, such as the mean and variance, do not only depend upon the spatial organization of raw data, as reflected for example in their spatial autocorrelation coefficient, but also on the aggregation process, namely on whether information is either averaged or summed.

To get insights from more realistic data configurations, let us extend this literature and compare the distortions arising from both a random and a sorted process of spatial assignment of simulated data. Consider a unit segment with $10,000$ equally-spaced addresses.[7] Each address is given the occurrence of a log-normally-distributed variable.[8] To study size distortions, we aggregate the addresses so as to form spatial units that constitute a partition of the unit-segment. First, we choose equally-shaped spatial units. Then, we consider randomly-shaped spatial units, that do not include the same number of addresses. To see whether size distortions depend on how information is aggregated, we study four polar cases: data summation or

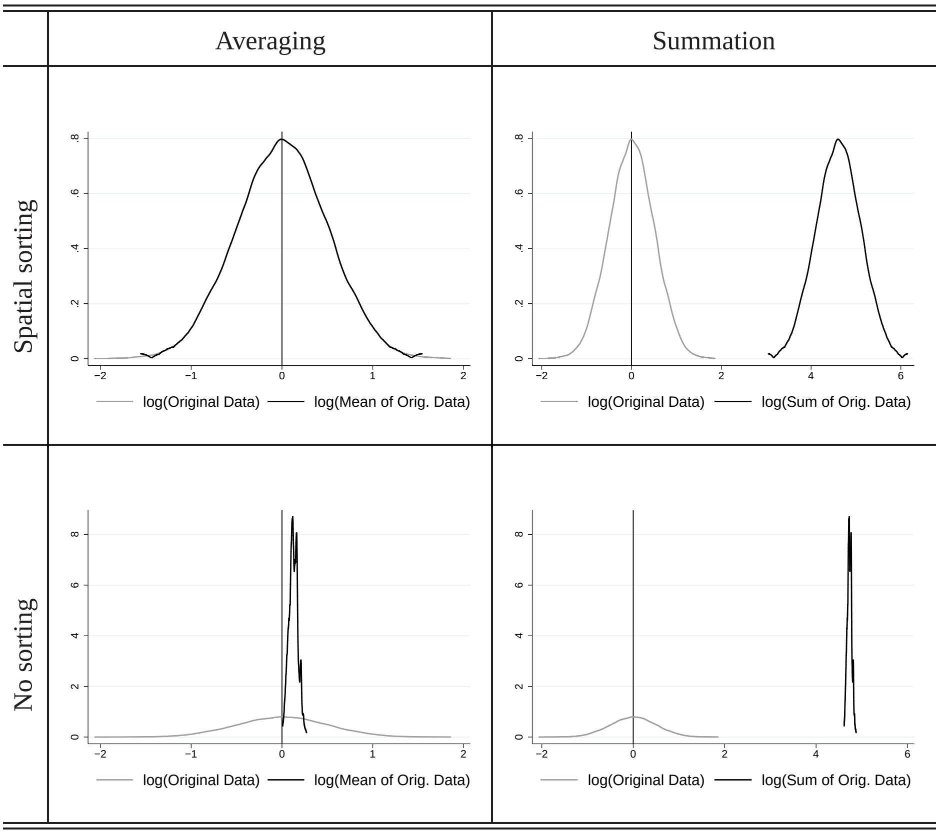
[5] More technically, Amrhein (1995) considers that the Cartesian coordinates of addresses are distributed either uniformly or normally, and that the generated variable follows either a normal or an uniform distribution.

[6] Under the same assumption of randomness, Holt, Steel, Tranmer, and Wrigley (1996) are able to justify theoretically the findings of Amrhein (1995). Note that Reynolds (1998) generates more realistic data configurations allowing for spatial autocorrelation.

[7] A two-dimension analysis of the MAUP would be more informative, but it is largely beyond the scope of the paper.

[8] The logarithm of the variable has a mean equal to 0 and a variance equal to 1.

Figure 2 – Aggregation with identically-shaped spatial units

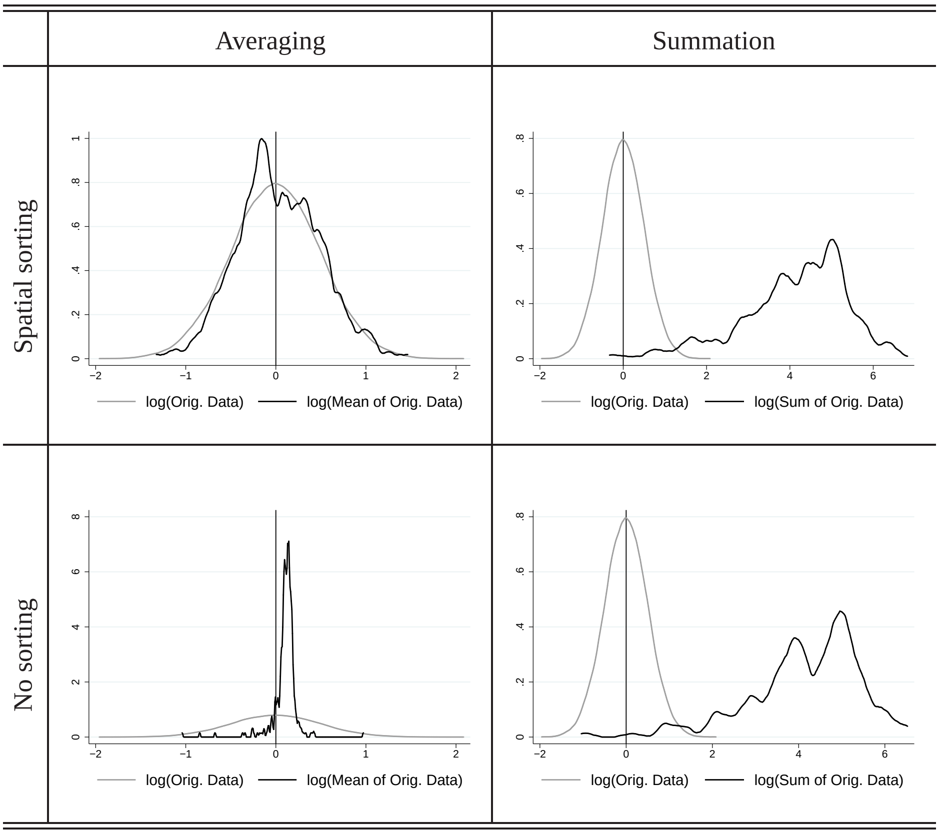


averaging over the addresses of each spatial unit, with either perfect or no data sorting over address values. In the unsorted configuration, the value at a given address is independent of surrounding addresses, as is the case in Amrhein (1995). In the perfectly sorted configuration, the addresses are ranked by increasing order of their assigned values before aggregation. Figure 2 compares the log-distribution of the simulated data (tight line) with their log-distribution when spatial units are equally-shaped (thick line).[9] Three main conclusions emerge:

1. Mean and variance can be almost perfectly recovered after aggregation when data are spatially sorted, regardless of whether data are averaged or summed (top graphs). The support of the distribution is only slightly reduced after aggregation. In the case of summation, the distribution is shifted to the right by a constant that depends on the number of aggregated addresses.
2. More information is lost when data are not sorted. While the mean is more or less correctly inferred after aggregation (up to the above constant), the variance is greatly reduced (by a ten-fold factor with

[9] We define units such that they include 100 contiguous addresses.

Figure 3 – Aggregation with randomly-shaped spatial units



our parametrization).

3. In any case, the distribution form remains more or less the same, and keeps its single peakness.

Subsequently, with low *within-unit* heterogeneity (e.g. spatial sorting) and low *between-unit* heterogeneity (e.g. identically-shaped units), the first moments of the distribution are not too much distorted by aggregation and changes in the size of units. By way of contrast, with strong *within-unit* heterogeneity (e.g. unsorted data), aggregation yields a loss of information, even if units are shaped homogeneously.

Figure 3 shows that aggregation is likely to raise more concerns when spatial units are randomly-shaped:

1. When data are both sorted and averaged (top left graph), information can be partially recovered.
2. This is not the case anymore when data are unsorted (bottom left graph). As before, the variance is drastically reduced.
3. Summation is more problematic with randomly-shaped units, even if data are perfectly sorted (top right graph): it does not only shift the distribution to the right (so as equally-shaped units), but it also enlarges the distribution support, thereby yielding an increasing dispersion of the variable.

To put it in a nutshell, when spatial units do not have the same shape, averaging is less sensitive to changes in size than summation, though part of the information is lost when data are not spatially-sorted. Conversely, if spatial units are randomly-shaped, summation is more distorted by a shift in their size. Distortions are even worse than data are unsorted.

2.3. Correlations distortions

Clear theoretical underpinnings are more difficult to come by for correlations, would they be univariate or multivariate. Fotheringham and Wong (1991), who consider a multivariate analysis of the determinants of mean household income for various zoning systems, come to an alarming conclusion: "The MAUP [...] is shown to produce highly unreliable results in the multivariate analysis of data drawn from areal units". They also find a sizable range for correlation and regression coefficients, which are positively (or negatively) significant for certain data configurations, but insignificant for others, suggesting that correlation inference is not robust to the aggregation process. Amrhein (1995) is the first to suggest separating aggregation effects from other types of discrepancies, such as model mis-specification in multivariate settings. In his simulation exercise, he shows that bivariate regression coefficients and Pearson correlations are sensitive to changes in the size and shape of spatial units, even if we know the data generation process and if we force the correlation between the two randomly-generated variables to be zero. However, he reaches a less alarming conclusion than Fotheringham and Wong (1991), and suggests that, for well-specified models, such as Amrhein and Flowerdew (1992), aggregation does not produce too many distortions, whereas for others, like Fotheringham and Wong (1991), the estimates are contaminated by size and shape.

Let us come back to our simulation exercise and turn to the analysis of regression coefficients. If aggregation distorts the explanatory and dependent variables in the same way, the size effect should be small. This is the case when, for instance, both the explanatory and dependent variables are spatially autocorrelated and averaged (top-left graph of figure 3). In sharp contrast, the size issue is more prevalent when the dependent and explanatory variables are not aggregated under the same process or do not exhibit the same degree of spatial autocorrelation.

As for shape distortion, it can be considered as a standard errors-in-variables issue. Let us consider the relationship $y^* = \beta_0 + \beta_1 x^* + \mu$, where y^* and x^* are two random variables, β_0 and β_1 two parameters, and μ an error term uncorrelated with x^*, and assume that the relationship is valid for a particular zoning system. Then, change the shape of spatial units so as to have $y = y^* + \epsilon$ and $x = x^* + e$ for the new spatial units. It is straightforward to show that, under this new zoning system, regressing $y = \widetilde{\beta}_0 + \widetilde{\beta}_1 x + \nu$ gives a biased estimator of β_1:

$$\hat{\tilde{\beta}}_1 = \beta_1 + \underbrace{\frac{cov(x^*,\epsilon) - \beta_1 cov(x^*,e) + cov(e,\epsilon) - \beta_1 \mathbb{V}[e]}{\mathbb{V}[x^*] + \mathbb{V}[e] + 2cov(x^*,e)}}_{bias}. \tag{1}$$

Note that there is no reason why the second right-hand term of equation (1) should be zero, except for knife-edge spatial configurations. Conversely, if the aggregation process generates random errors only and hence, $cov(x^*,e) = 0$, $cov(x^*,\epsilon) = 0$ and $cov(e,\epsilon) = 0$, the bias tends towards zero when $\mathbb{V}[x^*]$ grows faster than $\mathbb{V}[e]$. The larger the changes in borders, the larger the errors ϵ and e and thereby, the shape effect. Importantly, under the weaker condition that x is exogenous in the OLS regression of y on x, the bias is also zero.[10] In this respect, correcting the endogeneity of x, for instance with instrumental variables techniques, should alleviate the MAUP issue. Alternatively, improving specification should also reduce shape distortions, by making the explanatory variables more exogenous.

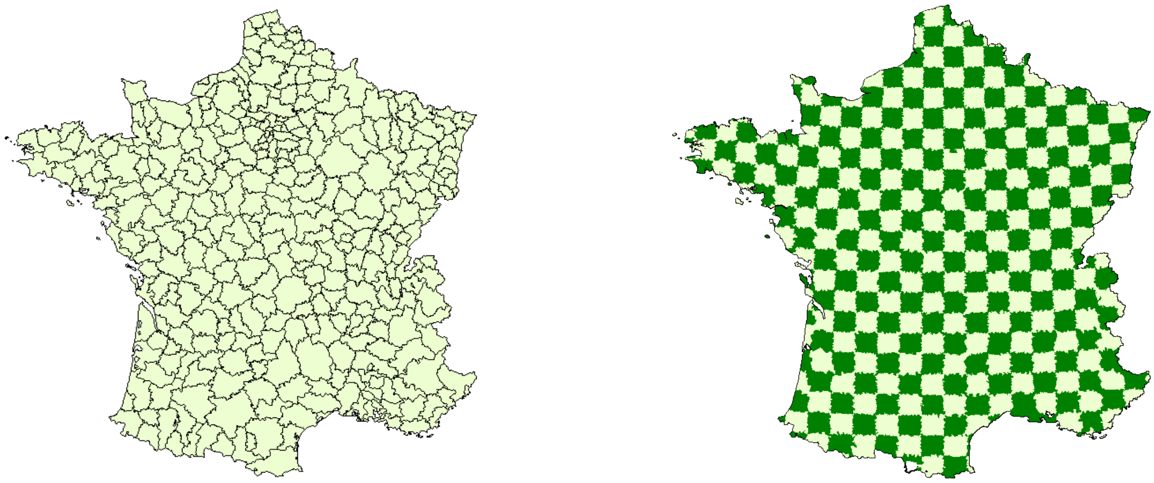
However, this exogeneity condition is not fulfilled if the value of x^* in one unit affects the outcome of the surrounding units (and therefore e, y^* and ϵ). The bias definitively increases with $cov(x^*,\epsilon)$ and $cov(e,\epsilon)$, i.e. with spatial correlation between x^* and y^*. By way of contrast, own spatial autocorrelation, reflected in $cov(x^*,\epsilon)$, has a mixed effect on the magnitude of the bias. This is due to the spatial sorting effect highlighted in section 2.2, which mitigates the negative impact of non-random errors.

In what follows, we build on these intuitions to extend the MAUP literature in a number of ways. First of all, we systematically assess the magnitude of size and shape distortions relative to mis-specification biases. Secondly, we examine different aggregation processes to test the sensitivity of economic inference to the MAUP. In wage-density regressions, raw information is averaged over spatial units, while for gravity regressions it is either summed or averaged. In light of the above discussion, the former should be associated with less distortions than the latter and thereby, the distribution of wages and density variables should be barely unmodified by changing zoning systems. In contrast, the trade dependent variable might well experience an enlargement of its distribution support, whereas the dispersion of most of the trade explanatory variables should shrink. Therefore, MAUP distortions should be more salient in gravity regressions. Finally, we extend the work of Fotheringham and Wong (1991) by comparing the estimates from six different administrative and grid zoning systems to those from a hundred equivalent random systems.[11]

[10] We have $y^* = \beta_0 + \beta_1 x^* + \mu \Rightarrow y = \beta_0 + \beta_1 x + \mu + \epsilon - \beta_1 e$. Variable x is exogenous if and only if $cov(x, \mu + \epsilon - \beta_1 e = 0) \Leftrightarrow cov(x^*,\epsilon) - \beta_1 cov(x^*,e) + cov(e,\epsilon) - \beta_1 \mathbb{V}[e] = 0$.

[11] In this respect, our study echoes the work of Holmes and Lee (forthcoming), who investigate the prevalence of a Zipf's law

Figure 4 – Small zoning systems



341 Employment Areas (EA) 341 Small squares (SS)

3. Zoning systems and data

The first zoning system we consider is that composed of 341 Mainland "Employment areas" (hereafter EA). These spatial units are underpinned by clear economic foundations, being defined by the French National Institute of Statistics and Economics (INSEE) so as to minimize daily cross-boundary commuting, or equivalently to maximize the coincidence between residential and working areas. This zoning system, currently composed of 341 areas, was designed to reduce the statistical artifact due to boundaries, which is why it is widely used in France. As can be seen on the left-hand side of figure 4, the average employment area is fairly small, covering 1,570 km^2, which is equivalent to splitting the U.S. continental territory into over 4,700 units.

Shape distortions can be identified from spatial units that are similar in size (or number) to employment areas. Conversely, size distortions can be highlighted with partitions of France involving units that are larger than the EAS. Hence, to disentangle the two faces of the MAUP, we appeal to three other sets of zoning systems.

for the U.S., based on an arbitrarily-drawn grid zoning system. It is also closely related to Menon (2008), who uses randomly-generated zoning systems equivalent to the commuting-defined Core Based Statistical Areas to study industrial agglomeration in the US.

3.1. Administrative zoning systems

The first set refers to French administrative units. Continental France is partitioned into 21 administrative "Régions" (RE), depicted on the left of figure 5, which are themselves split into 94 "Départements" (DE), shown on the left of figure 6. All such units are aggregates of municipalities, the finest spatial division for which data are available in France.[12]

It can nonetheless be argued that administrative boundaries do not capture the essence of economic phenomena that often spill over boundaries, which is one of the reasons why EAs were created. To circumvent this drawback, some authors, especially geographers, prefer to work with (often arbitrarily-drawn) checkerboard grids. The rationale is that, even if they do not necessarily better match the "true" boundaries of economic phenomena, grid zoning systems provide a greater degree of spatial homogeneity than do administrative zoning systems.[13]

3.2. Grid zoning systems

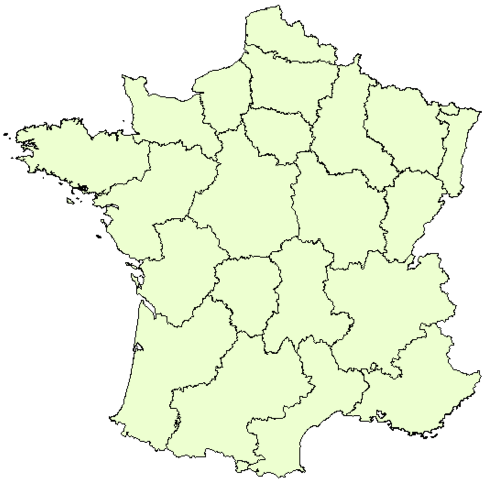
We therefore construct a second set of zoning systems purely based on grid units. We first enclose France into the smallest possible rectangle. We then divide this rectangle into lattices of squares (based on longitude and latitude). As France is more or less hexagonal, several squares jut out into the sea and we obviously left this out. We obtain the final grid by aggregating all municipalities which have their centroid into the same square. The resulting units are not perfect squares as their boundaries follow those of real municipalities. We choose the size of the squares to produce three different zoning systems analogous to administrative ones: 22 (non-empty) large squares (LS), 91 medium squares (MS) and 341 small squares (SS). It is worth noting that the largest zoning systems (LS and MS in figures 5 and 6) include several squares which are partially truncated due to French national boundaries. The finest grid such as SS (figure 4) circumvents this pitfall at the expense of geometry, since the units boundaries become increasingly ragged at the very fine scale. Therefore, overtly enlarging or tightening the units alters both their symmetry and regularity.

A comparison of the results obtained under respectively RE, DE and EA or LS, MS and SS gives a flavor of any size distortions. We capture the impact of shape by comparing the results obtained across zoning systems involving units of similar size (RE to LS, DE to MS, and EA to SS). While these comparisons tell us

[12] The French metropolitan area is covered by 36,247 municipalities.

[13] Another argument is that grid zoning systems do not change over time, while administrative areas may do so. See ESPON (2006) for an overview of this issue.

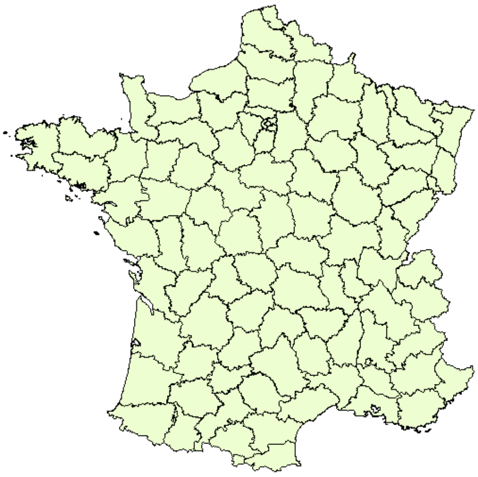
Figure 5 – Large zoning systems



21 Régions (RE)

22 Large squares (LS)

Figure 6 – Medium zoning systems



94 Départements (DE)

91 Medium squares (MS)

whether MAUP distortions exist, they do not indicate whether the differences in the results are systematic and significant, however, which is why we propose a third set of zoning systems.

3.3. *Partly random zoning systems*

Our third set of zoning systems involves arbitrarily-drawn spatial units. We define a set of 100 different partitions of France, by randomly aggregating the 4,662 French "Cantons",[14] into zoning systems that have a number of units strictly equivalent to those of administrative ones (341 units for EA, 94 for DE and 21 for RE): we call these REA, RDE and RRE respectively. These are constructed using the following algorithm. We randomly draw one canton, called the seed, within each administrative unit. We then aggregate each seed to a second canton randomly drawn from those contiguous to it. We continue with a third canton and so on, until all existing cantons have been drawn. We run the algorithm 100 times at each scale. Broadly speaking, this procedure produces, for each scale, a partition of France with jiggling borders.

3.4. *Characteristics of zoning systems*

Our empirical analysis builds on sectoral time-series data at the municipal level. The aggregation into the aforementioned larger zoning systems yields a three-dimension panel of employment, number of plants and wages for 18 years (within the 1976-1996 period) and 98 industries (at the two-digit level for both manufacturing and services). For 1996, we match this panel to a trade data set for manufactured goods.[15]

As can be seen in table 1, zoning systems differ sharply in their economic features. The spatial variation in land area is smaller for small grid units than for employment areas, a property that does not hold for larger administrative units. This reflects two opposite effects. On the one hand, grid units are more regular, which reduces the variance. On the other hand, the share of truncated grid units increases with size, which increases the variance. The latter effect dominates for medium and large units. A clear drawback of the grid strategy is that, when units are not small enough, the gains of reducing the variance of land area cannot be attained due to the irregularity of national borders. Conversely, this also shows that the French authorities were fairly successful in designing quite homogeneous administrative units.

Regarding the other variables, an important distinction concerns the way in which information is aggregated. Some variables, such as employment and trade flows, are *summed*, whereas others, such as job

[14] We use this intermediate grouping of French municipalities to reduce the computational time without losing too much spatial variability in the randomization process.

[15] More details on the data are provided in Appendix.

Table 1 – Summary statistics

Zoning system		(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Number of units		341	341	94	91	21	22
Land Area (km^2)	Av.	1569.8	1580.4	5733.3	5922.3	25663.4	24496.7
	Cv.	0.63	0.35	0.34	0.5	0.43	0.53
Employment (workers)	Av.	2012	2019	7300	7541	32678	31193
	Cv.	2.45	3.73	1.28	2.37	1.16	1.33
Employment density (workers/km^2)	Av.	4.6	1.5	12.3	1.7	1.8	1.3
	Cv.	8.7	3.1	6.3	1.7	1.8	0.8
Aggregate Market Potential	Av.	2910	2432	2956	2161	2137	1791
	Cv.	0.6	0.3	0.8	0.3	0.3	0.2
Municipality-level Market Potential	Av.	3300	2758	3585	2705	3097	2736
	Cv.	0.7	0.4	0.9	0.4	0.5	0.3
(Gross) Wage	Av.	1.3	1.2	1.3	1.3	1.3	1.3
	Cv.	0.2	0.2	0.1	0.1	0.1	0.1
Aggregate Distance (km)	Av.	393	419	384	445	371	448
	Cv.	0.49	0.48	0.5	0.49	0.52	0.52
Municipality-level Distance (km)	Av.	394	417	386	435	381	420
	Cv.	0.49	0.48	0.49	0.49	0.49	0.49
Trade Flow (tons $\times$ 1000)	Av.	91.05	102.22	382.83	496.8	5956.48	5839.38
	Cv.	6.6	6.5	5.6	6.2	3.1	3.2

Notes: (*i*) (EA): employment areas, (SS): small squares, (DE): Départements, (MS): medium squares, (RE): Régions, (LS): large squares. (*ii*) Averages over 18 years, except for trade flows (1996 value). (*iii*) Av. is the mean. Cv is the Coefficient of variation (standard deviation divided by mean). (*iv*) No unit for wage because detrended and centered around individual mean. No unit for market potential.

density and wages, are *averaged*. The former increase with the size of the units, which is straightforward. By way of contrast, the overall picture vary less for averaged information. For instance, employment density differs only little across grid zoning systems, regardless of the size of their units, while it varies more for administrative units, which reflects that the design of administrative zoning systems was not based on this variable. Average wages are little affected by both administrative and grid zoning systems. The suspicion that the MAUP could still bias the estimate of the impact of agglomeration economies motivates the exercise carried out in section 5.

However, there are two variables, distance and market potential, for which information is neither summed nor averaged. Consider first distance. It can be computed either as the great-circle distance between the centroids of spatial units ("Aggregate Distance" in table 1), or as the average distance between the municipalities of each unit ("Municipality-level Distance" in table 1). In the former case, there is no obvious link from one zoning system to the other, whereas in the latter, less information is lost through aggregation. The same argument holds for market potential. It can be the average of market potentials over municipalities or the aggregate market potential. Even if the two first moments of both couples of variables do not differ drastically, the MAUP could be more severe when variables are computed at the aggregate level. This source of distortions is investigated in sections 5 and 6.

4. Spatial concentration

Before turning to regression analysis, we carry out the most basic exercise in economic geography, which consists in measuring the extent of spatial concentration, an issue widely-covered in the literature. Apart from a small number of continuous approaches, such as Duranton and Overman (2005), work in this area is based on discrete zoning systems. While some work has focused on the comparison of spatial concentration across industries, such as Ellison and Glaeser (1997), only little has assessed the legitimacy of comparing results across zoning systems that differ in the size and shape of spatial units. In this section, we compare the variability in concentration due to the zoning system with that from different concentration indices.

4.1. Gini indices

We compute the spatial Gini index associated with every zoning system for 98 industries and 18 years (see Appendix). The moments of the index distribution are provided in table 2. Every moment of the distribution, in particular the mean, falls with aggregation level. The rationale is straightforward: smallers units have more areas with no registered employment for certain industries, which raises the Gini index mechanically for each industry.

We then rank industries by spatial concentration and compute Spearman rank correlations across zoning systems. The results are shown in table 3.

Rank correlations across zoning systems that are similar in size (EA and SS, DE and MS, and RE and LS) are very high, with values of at least 0.98 (see the sub-diagonal elements in table 3). The ranking of

Table 2 – Summary statistics for the Gini index

	Mean	St. Dev.	Min	P25	P50	P75	Max
(ZE)	0.587	0.224	0.134	0.410	0.597	0.767	0.994
(SS)	0.553	0.220	0.111	0.370	0.560	0.720	0.992
(DE)	0.481	0.217	0.098	0.299	0.465	0.637	0.980
(MS)	0.439	0.213	0.072	0.260	0.415	0.582	0.971
(RE)	0.338	0.187	0.051	0.184	0.321	0.443	0.947
(LS)	0.327	0.185	0.043	0.181	0.300	0.433	0.891

Note: Computed on 1764 observations (98 industries × 18 years).

Table 3 – Spearman rank correlations between Gini indices

Averages over 18 years

	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
(EA)	1	0.99	0.99	0.99	0.95	0.95
(SS)		1	0.98	0.99	0.96	0.96
(DE)			1	0.99	0.97	0.97
(MS)				1	0.98	0.98
(RE)					1	0.98
(LS)						1

industries is therefore virtually unaffected by changes in the shape of units. Size has a slightly greater effect on concentration. For instance, the rank correlation between EA and RE is 0.95, which remains high. Making shape more homogeneous across scales leads to similar results, with the correlation between SS and LS zoning systems being 0.96.

4.2. *Ellison and Glaeser indices*

It is well known that the spatial Gini index is contaminated by industry structure. Given total industry employment, industries with fewer plants will have higher Ginis, even with random plant location. Ellison and Glaeser (1997) develop a measure of concentration that is purged of this plant size effect. Table 4 describes moments of the EG index distribution.

Contrary to the Gini coefficient, the EG index monotonically increases with the aggregation scale, which gives further support to well-known result already put forward by Ellison and Glaeser (1997), or Maurel and Sédillot (1999) and Devereux, Griffith, and Simpson (2004), for a slightly modified index. It

Table 4 – Summary statistics for the Ellison-Glaeser index

	Mean	St. Dev.	Min	P25	P50	P75	Max
(ZE)	0.017	0.027	-0.015	0.004	0.009	0.019	0.396
(SS)	0.021	0.037	-0.065	0.004	0.012	0.027	0.365
(DE)	0.022	0.034	-0.014	0.005	0.012	0.025	0.407
(MS)	0.031	0.051	-0.067	0.004	0.014	0.039	0.364
(RE)	0.042	0.059	-0.062	0.006	0.023	0.051	0.434
(LS)	0.040	0.056	-0.116	0.005	0.018	0.052	0.326

Note: Computed on 1764 observations (98 industries × 18 years).

can be taken as evidence that various industrial spillovers play at different scales. If we turn to the Spearman rank correlations, we have the results depicted in table 5.

Table 5 – Spearman correlations between EG indices

Averages over 18 years

	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
(EA)	1	0.83	0.94	0.84	0.83	0.81
(SS)		1	0.79	0.87	0.85	0.84
(DE)			1	0.85	0.85	0.82
(MS)				1	0.93	0.90
(RE)					1	0.94
(LS)						1

The rank correlations are generally lower than those for the Gini indices. Hence, any distortions due to the MAUP are more pronounced when spatial concentration is measured via the EG index. In particular, size distortions are slightly aggravated, even though the rank correlations remain fairly high (0.83 for instance between EA and RE).

4.3. *Comparison between the Gini and the EG*

The success of the EG index over the Gini coefficient lies in its alleviation of concentration due to the location of big plants. In this respect, the EG index should be favored. The crucial question we address here is whether the zoning system affects the ranking of industries more than does the choice of the index itself. To answer, we turn to a between-index rank correlation analysis.

Table 6 shows that the between-index Spearman rank correlations are definitely smaller than their within counterparts. Even within each zoning system (the diagonal elements of table 6), the rank correlation is 0.81 at best (for RE), with the lowest correlation being 0.56 (for SS).

Table 6 – Spearman rank correlations between Gini and EG indices

Averages over 18 years

	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
(EA)	0.65	0.53	0.70	0.61	0.67	0.64
(SS)	0.65	0.56	0.70	0.63	0.69	0.66
(DE)	0.69	0.56	0.75	0.65	0.71	0.67
(MS)	0.68	0.58	0.74	0.67	0.73	0.69
(RE)	0.73	0.65	0.78	0.74	0.81	0.76
(LS)	0.73	0.65	0.78	0.73	0.79	0.78

There is considerable evidence that index choice, which we can consider as a specification issue, produces greater distortions than the choice of zoning system, in terms of both size or shape. It should thus be of greater concern than the MAUP.

5. Agglomeration economies

While the MAUP only slightly distorts spatial concentration patterns, it might have a greater effect on the explanation of the spatial distribution of economic variables. We therefore now consider the incidence of the MAUP in the context of multivariate regression analysis. In this section, we focus on the estimation of agglomeration economies. Evaluating the magnitude of the benefits reaped from spatial proximity is important for policy, and much work, such as Ciccone and Hall (1996), has been devoted to the estimation of the productivity gains resulting from dense clusters of activities. The benefits from proximity to large markets and the local composition of labor skills are generally simultaneously estimated.[16]

We regress local wages, a frequently-used measure of local labor productivity,on local employment density. Let w_{at} denote the wage in area a at date t, computed as the *average* earnings of all workers located in a at date t (hereafter the "gross" wage), and Den_{at} employment density (per square-kilometer). The benchmark specification we run is the following:

$$\log w_{at} = \alpha \log Den_{at} + \gamma X_{at} + \varepsilon_{at}, \tag{2}$$

[16] See for instance Combes, Duranton, and Gobillon (2008a).

where X_{at} is a vector of control variables. We compare the estimated elasticity of wages to employment density across zoning systems. In this exercise, we consider the average wage and employment density per areal unit. In light of the simulations performed in section 2, we expect the MAUP to be mitigated in this setting. As for concentration indices, we then check whether the choice of zoning systems matters less for the magnitude of agglomeration economies than the biases from choice of controls in the wage equation, which is a specification issue.

5.1. *A wage - density simple correlation*

In order to have a benchmark, we first look at gross wage / density correlations. Given the panel structure of the data, we estimate equation 2 with no controls other than time dummies. Table 7 reports on the resulting elasticities.

Table 7 – Gross wages and density

Simple correlations

	Dependent Variable: Log of gross wage (pooled years)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Density	0.071^a	0.070^a	0.073^a	0.050^a	0.090^a	0.099^a
	(0.001)	(0.002)	(0.001)	(0.002)	(0.003)	(0.006)
Time dummies	yes	yes	yes	yes	yes	yes
Obs.	6138	6118	1692	1638	378	396
R^2	0.468	0.237	0.729	0.376	0.762	0.549

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

The elasticity of wages with respect to employment density lies in the usual range of [0.04, 0.10] found for U.S. and European data (see Combes et al., 2008b). Even though some differences result from the move to a larger scale, the shape effect remains small.

Size differences do not really matter when moving from small to medium units, although larger differences occur as we move to the largest units. In both EA and DE, the value is about 0.07. However, the aggregation from DE to RE induces a 20%-increase in the coefficient estimate. As for the grid zoning system, the estimated elasticity is more sensitive to scale.

It is worth noting that the explanatory power of employment density is significantly lower (almost halved) for checkerboard grids than for administrative units. Therefore, boundaries which do not reflect administrative/economic realities do actually generate measurement errors, possibly in both the left-hand and right-hand side variables. However, the good news is that these errors seem to be largely randomly distributed: even though density loses explanatory power, the overall picture with respect to elasticity is one of stability. In line with the intuitions provided in 2.3, this corroborates the OLS consistency in the presence of random measurement errors and exogenous explanatory variables.

As a second step, we compare the two MAUP distortions to the changes induced by including skills controls (Section 5.2) and market potential (Section 5.3) into the wage equation.

5.2. Controlling for skills and experience

Our empirical analysis uses rich individual wage information from a large panel of workers followed across time and jobs. We are hence able to apply a sophisticated procedure to control for observed and unobserved individual skills, so as to check whether the greater productivity observed in dense areas is partly due to the spatial sorting of workers and whether the MAUP affects these magnitudes. In a first stage, we calculate individual wages net of individual skills and experience, as follows:

$$\log w_{it} = \theta_i + \nu_{j(i,t)} + X_{it}\beta + \epsilon_{it}, \tag{3}$$

where w_{it} is the wage of worker i at date t. This is a function of θ_i, an individual fixed-effect capturing the impact of both time-invariant observed and unobserved skills, $\nu_{j(i,t)}$, an effect specific to the firm j where i is employed at date t, and X_{it} a set of controls for worker's i experience at date t (age, age-squared, and number of previous jobs interacted with gender). Based on the estimates provided in Abowd, Creecy, and Kramarz (2002), and following Combes et al. (2008a), we define a wage net of any individual observed and unobserved skills and experience effects, $\left(w_{it} - \hat{\theta}_i - X_{it}\hat{\beta}\right)$. We then compute the average of this net wage over all individuals living in the same area a, at date t (hereafter net wage). This yields a measure of local labor productivity purged of individual skills and experience. We proceed by regressing net wages on employment density. The results are shown in table 8.

The elasticity of net wages with respect to employment density is half of that for gross wages. Hence, the specification issue induces a difference in coefficient of an order of magnitude greater than that due to the MAUP. We therefore reach the same conclusion as for the analysis of spatial concentration: differences due to the size and shape of spatial units are small compared to the upward bias induced by the omission of workers' skills and experience in the wage equation, especially when data are not aggregated at a too large

Table 8 – Net wages and density

Simple correlations

	Dependent Variable: Log of net wages (pooled years)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Density	0.033^a	0.028^a	0.029^a	0.023^a	0.048^a	0.052^a
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.004)
Time dummies	yes	yes	yes	yes	yes	yes
Obs.	6138	6118	1692	1638	378	396
R^2	0.220	0.098	0.338	0.238	0.619	0.570

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

scale. Moreover, shape and size distortions are slightly attenuated in many cases (between DE and MS, and RE and LS, for instance), once these controls are included.

5.3. *Market potential as a new control*

Not only local density and skill composition affect labor performance, but so does the proximity to large economic centers outside the area. A major drawback of the above wage specifications is that there are no controls for the relative position of the area within the whole economy. For instance, wage equations derived from fully-specified economic geography models, such as Redding and Venables (2004) and Hanson (2005), account for spatial proximity via structural demand and supply access variables. It is beyond the scope of this paper to replicate such a sophisticated and difficult to implement approach. Here we only include, as well as density, a Harris (1954) market potential variable based on the employment accessible from any given area, divided by the distance necessary to reach them:[17]

$$\text{Market Potential} = \sum_{a' \neq a} \frac{Y_{a'}}{Dist_{a,a'}}, \tag{4}$$

where $Y_{a'}$ is employment in area a and $Dist_{a,a'}$, the great-circle distance between the centroids of areas a and a'. The results for gross and net wages are listed in tables 9 and 10 respectively.

[17]The literature shows that this atheoretic market potential often has an explanatory power similar as the one of structural market potential.

Table 9 – The spatial determinants of gross wages

	Dependent Variable: Log of gross wage (pooled years)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Density	0.055^a	0.065^a	0.059^a	0.050^a	0.090^a	0.098^a
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.006)
Market Potential	0.100^a	0.099^a	0.062^a	0.079^a	0.024^b	-0.009
	(0.004)	(0.008)	(0.005)	(0.008)	(0.011)	(0.020)
Obs.	6138	6118	1692	1638	378	396
R^2	0.521	0.256	0.753	0.411	0.765	0.549

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

Once market potential is accounted for, the impact of density on gross wage is attenuated. This is even more salient for low-scale and administrative zoning systems. The elasticity of gross wages to market potential is slightly stronger for medium squares than for their administrative counterparts, Départements. This is consistent with the intuition that cross-boundary discrepancies should be more salient for grid units that were not designed to minimize them in the first place.

Regarding the size issue, the impact of market potential monotonically decreases with the aggregation scale (for both the administrative and grid zoning systems). As for density, size distortions are more prevalent for RE or LS, and market potential becomes either insignificant or even negative. This is due to an important loss of information in the aggregation process, that we detail below.

In table 10 where skill controls are accounted for, shape and size alter only slightly the estimates at the lowest scales. It confirms our previous result that specification is of primary concern when working with small spatial units. Differences due to size and shape are much less pronounced than those resulting from a change in specification. For instance, the elasticity of density is only 0.027 at the small-unit levels, once skills and market potential are controlled for, while the baseline estimates were about 0.07. Similar conclusions are reached for the market potential elasticities, with slightly larger differences at the largest scales (RE and LS). To gain further insights on the underpinnings of such large-scale discrepancies, we turn to an alternative definition of market potential.

Table 10 – The spatial determinants of net wages

	Dependent Variable: Log of net wage (pooled years)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Density	0.027^a	0.026^a	0.021^a	0.023^a	0.048^a	0.052^a
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.004)
Market Potential	0.037^a	0.043^a	0.036^a	0.044^a	0.023^b	-0.0002
	(0.004)	(0.007)	(0.006)	(0.007)	(0.01)	(0.012)
Obs.	6138	6118	1692	1638	378	396
R^2	0.232	0.104	0.354	0.256	0.624	0.570

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

5.4. An alternative definition of market potential

If we use the average of municipality-level market potentials instead of the aggregate market potential, we obtain the results reported in tables 11 and 12.

Table 11 – The spatial determinants of gross wages:

Municipality-level market potential

	Dependent Variable: Log of gross wage (pooled years)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Density	0.050^a	0.061^a	0.051^a	0.038^a	0.063^a	0.069^a
	(0.001)	(0.002)	(0.002)	(0.002)	(0.004)	(0.006)
Market Potential	0.101^a	0.094^a	0.077^a	0.120^a	0.091^a	0.125^a
	(0.004)	(0.008)	(0.005)	(0.007)	(0.01)	(0.014)
Obs.	6138	6118	1692	1638	378	396
R^2	0.520	0.254	0.761	0.464	0.808	0.624

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

In this second set-up, the aggregation process conserves more information and, as expected, the elasticity of market potential is less sensitive to changes in the shape and size of units, and even less at the largest scales. Interestingly, the MAUP is also less salient regarding employment density, and the explanatory power of the model increases, in comparison with tables 9 and 10.

Table 12 – The spatial determinants of net wages:

Municipality-level market potential

	Dependent Variable: Log of net wage (pooled years)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Density	0.025^a	0.024^a	0.017^a	0.016^a	0.032^a	0.038^a
	(0.001)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
Market Potential	0.038^a	0.044^a	0.042^a	0.063^a	0.056^a	0.059^a
	(0.004)	(0.006)	(0.006)	(0.007)	(0.009)	(0.009)
Obs.	6138	6118	1692	1638	378	396
R^2	0.232	0.105	0.357	0.278	0.652	0.611

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

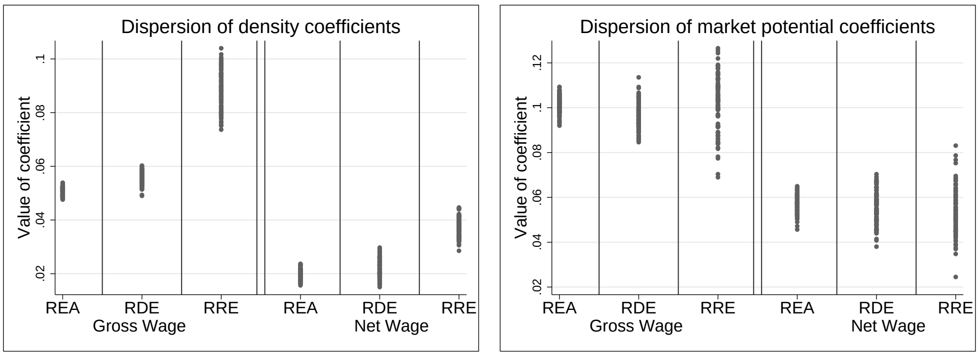
As for net wages (see table 12), the coefficients of both density and market potential are more than halved compared to gross wages, whereas shape and size are clearly not big issues.

Figure 7, that displays the density and market potential estimates drawn from the three partly random zoning systems, provides further support to this conclusion. For a given size, the dispersion of estimates is much lower than that induced by a shift of specification, which confirms the absence of shape effects. Once again, the only significant difference due to size regards density for the largest units. Even so, this distortion almost vanishes in the best specification (net wages), as do the differences in the impact of market potential. These conclusions clearly echo the findings of Amrhein and Flowerdew (1992) and suggest that a good specification is actually an efficient way to circumvent the MAUP.

In line with the simulations provided in section 2.3, the loss of information incurred when variables are aggregated is the primary source of the MAUP. It can be mitigated (but never completely eliminated) when the process of aggregation is of the average-type and when the raw information is not too much heterogeneous *within-unit*, which is the case for spatially autocorrelated data at small scales. If so, the MAUP is of secondary concern compared to modeling issues.[18]

[18] One important concern is not tackled here. In the above wage-density analysis, we inevitably face the major difficulty that causality could run both ways since the worker's location is also determined by their earnings anticipations. We leave this issue aside, as it has already been extensively discussed in the literature, and is orthogonal to the MAUP.

Figure 7 – The size- and shape-dependency of wage determinants



Note: (REA): Random employment areas, (RDE): Random Départements, (RE): Random Régions.

6. Gravity equations

So far, we have investigated MAUP distortions for aggregations processes that are of the average-type only. We now turn to gravity regressions that need both averaged and summed information.

6.1. Basic gravity

The gravity model has been widely used to investigate the determinants of trade. A basic specification explains the trade flow $F_{aa'}$, originating from area a and shipped to area a', by various proxies for the proximity between a and a'. These include the great-circle distance between the centroids of a and a', $Dist_{aa'}$ and, often, a dummy variable stating whether the areas are contiguous, $Contig_{aa'}$.[19] Finally, the "border effect" (see McCallum, 1995) is captured by a dummy variable for within-area flows, $Within_{a=a'}$. As a first step, we estimate the following two-way fixed-effect specification:

$$\ln(F_{aa'}) = \theta_a + \theta_{a'} - \rho \ln(Dist_{aa'}) + \phi Contig_{aa'} + \psi Within_{a=a'} + \epsilon_{aa'}, \qquad (5)$$

where θ_a and θ'_a are destination and origin fixed effects respectively, and $\epsilon_{aa'}$ is an error term. This fixed-effect approach has the attractive property of being structurally compatible with many trade models (based on comparative advantage as well as imperfect competition).[20]

Table 13 reports on the related estimates under both the administrative and grid zoning systems. The great-circle distance elasticity is systematically larger for grid than for administrative zoning systems, at a

[19] A for grid zoning systems, we assume that two units are contiguous if they share a common edge.
[20] See Feenstra (2003).

Table 13 – Basic gravity

Aggregate distance

	Dependent Variable: log of positive flows (Year 1996)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Distance	-0.996[a]	-1.175[a]	-1.608[a]	-1.912[a]	-1.602[a]	-1.900[a]
	(0.022)	(0.024)	(0.056)	(0.048)	(0.075)	(0.113)
Within	1.738[a]	1.040[a]	1.395[a]	0.221	1.460[a]	0.445[b]
	(0.063)	(0.066)	(0.111)	(0.135)	(0.151)	(0.211)
Contiguity	0.967[a]	1.093[a]	0.959[a]	1.044[a]	0.728[a]	0.895[a]
	(0.041)	(0.044)	(0.063)	(0.077)	(0.087)	(0.118)
Obs.	24849	22189	6600	5069	441	443
R^2	0.516	0.541	0.706	0.752	0.941	0.928

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) [a], [b], [c]: Significant at the 1%, 5% and 10% levels respectively.

given scale. The shape effect on distance increases with the scale of aggregation. Contiguity is less affected by shape. Again, size effects are slightly more salient at the largest scales, especially when moving from the EA-SS to either the DE-MS or RE-LS zoning systems. The magnitude of the distance effect (in absolute value) increases with size (for the administrative and grid zoning systems). The border effect is always lower for grid zoning systems, which is further evidence of the economic consistency of administrative units.

If we use the average of inter-municipality distance instead of aggregate distance (see table 14), results remain virtually the same, but the border effect is magnified.

In sharp contrast with market potential in wage equations, an alternative measure of distance does not alleviate the MAUP. Gravity regressions are hence more sensitive to the MAUP. The rationale is found in the simulations depicted in section 2.3. The dependent variable, trade flows, is summed over units, whereas the explanatory variable, distance, is averaged. The process of aggregation shifts to the right the distribution of the former and raises its dispersion (which finds support in table 1). By way of contrast, since distance is a highly autocorrelated averaged variable, it is less sensitive to aggregation. The rise (in absolute value) of the distance coefficient reflects the need to reconciliate an increasing dispersion of trade flows with a stable support of the distance distribution.

Figure 8 illustrates the way in which both size and shape affect the values and standard errors of esti-

Table 14 – Basic gravity

Municipality-level distance

	Dependent Variable: log of positive flows (Year 1996)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Weighted Distance	-1.009[a]	-1.182[a]	-1.645[a]	-1.909[a]	-1.710[a]	-1.968[a]
	(0.022)	(0.024)	(0.058)	(0.047)	(0.088)	(0.096)
Within	2.139[a]	1.547[a]	1.938[a]	1.138[a]	1.900[a]	1.395[a]
	(0.058)	(0.056)	(0.099)	(0.097)	(0.146)	(0.21)
Contiguity	1.031[a]	1.139[a]	1.020[a]	1.058[a]	0.768[a]	0.863[a]
	(0.04)	(0.044)	(0.062)	(0.069)	(0.082)	(0.094)
Obs.	24849	22189	6600	5069	441	443
R^2	0.517	0.544	0.709	0.757	0.942	0.933

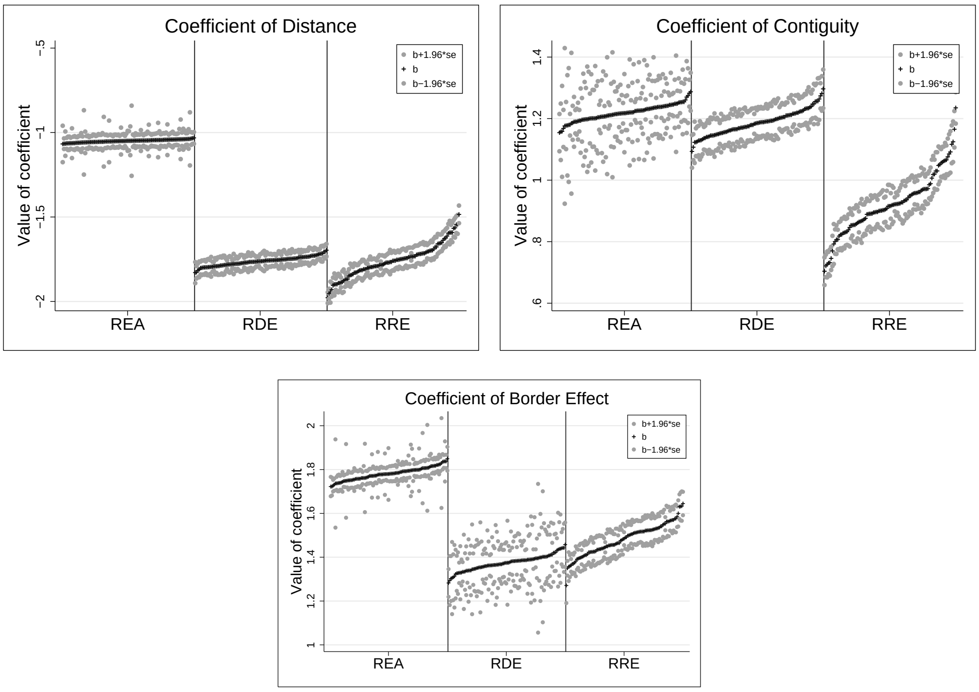
Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) [a], [b], [c]: Significant at the 1%, 5% and 10% levels respectively.

mates from partly random zoning systems. Dark dots in the top-left figure stand for the elasticity of distance (and for contiguity and border effects in the top-right and bottom figures respectively). The 95% confidence interval is shown by the surrounding lighter dots. Random zoning systems are ranked by increasing estimated values. For all three proximity measures, we find that the variability in estimates raises with scale (as reflected by the increasing slope of dark curves), suggesting more shape-dependency in larger zoning systems. Nonetheless, this variability is of lower magnitude than the differences due to moving from one scale to another (from REA to RDE or RRE, regarding distance and border effects). The shape-dependency of larger zoning systems (especially RRE) is due to two joint phenomena. First, coefficient estimation is more likely to suffer from finite-sample bias for larger (and hence less numerous) units. Second, the random process of aggregation is likely to produce more distinct zoning systems when data are aggregated over larger units.

6.2. *Augmented Gravity*

Barriers to trade do not only concern proximity. Other trade frictions result from costs unrelated to distance (such as trade policy, exchange-rate volatility, delivery times, and inventory or regulation costs), and from more subtle frictions due to the need to acquire information on remote trading partners or to enforce contracts, as emphasized by Rauch (2001). To tackle these, the literature extends the basic gravity model

Figure 8 – The size- and shape-dependency of the impact of spatial proximity on trade



Note: (*i*) The coefficients (*b*) have to be greater (in absolute value) than 1.96 times the standard error (se) to enter into the 95% confidence interval. (*ii*) (REA): Random employment areas, (RDE): Random Départements, (RRE): Random Régions.

by making trade costs depend not only on spatial proximity but also on cultural and informational proximity. For instance Wagner, Head, and Ries (2002) report that migrations between two countries enhance their bilateral trade by around 50%. To evaluate the trade-creating impact of social and business networks within countries, Combes, Lafourcade, and Mayer (2005) estimate:

$$\begin{aligned} \ln(F_{aa'}) &= \theta_a + \theta_{a'} - \rho \ln(Dist_{aa'}) + \phi Contig_{aa'} + \psi Within_{a=a'} \\ &+ \alpha \ln(1 + Mig_{aa'}) + \beta \ln(1 + Mig_{a'a}) + \gamma \ln(1 + Plant_{aa'}) + \epsilon_{aa'}, \end{aligned} \tag{6}$$

where $Dist_{aa'}$ is municipality-level distance,[21] $Mig_{aa'}$ is the number of people born in area a' and working in area a, called (relative to area a) immigrants, $Mig_{a'a}$ are analogously emigrants, and $Plant_{aa'}$ is the number of financial connections between plants belonging to the same business group (see Appendix).

Table 15 – Augmented Gravity

	Dependent Variable: log of positive flows (1996, Municipality-level distance)					
Zoning system	(EA)	(SS)	(DE)	(MS)	(RE)	(LS)
Distance	-0.616^a	-0.698^a	-1.231^a	-1.294^a	-1.291^a	-1.340^a
	(0.023)	(0.027)	(0.062)	(0.061)	(0.103)	(0.102)
Within	1.201^a	0.925^a	0.8^a	0.338^a	0.517^a	0.436^c
	(0.064)	(0.06)	(0.126)	(0.095)	(0.171)	(0.241)
Contiguity	0.315^a	0.403^a	0.366^a	0.317^a	0.296^a	0.425^a
	(0.049)	(0.049)	(0.068)	(0.072)	(0.072)	(0.119)
Emigrants	0.228^a	0.226^a	0.237^a	0.244^a	0.281^a	0.246^b
	(0.014)	(0.013)	(0.028)	(0.034)	(0.088)	(0.104)
Immigrants	0.241^a	0.256^a	0.209^a	0.286^a	0.257^a	0.268^b
	(0.014)	(0.015)	(0.037)	(0.035)	(0.086)	(0.134)
Business networks	0.043^a	0.013	0.24^a	-0.021	0.225	0.646^a
	(0.016)	(0.019)	(0.072)	(0.064)	(0.173)	(0.161)
Obs.	24849	22189	6600	5069	441	443
R^2	0.538	0.568	0.723	0.772	0.953	0.945

Notes: (*i*) All variables in logarithms. (*ii*) Standard-errors in brackets. (*iii*) ^{a}, ^{b}, ^{c}: Significant at the 1%, 5% and 10% levels respectively.

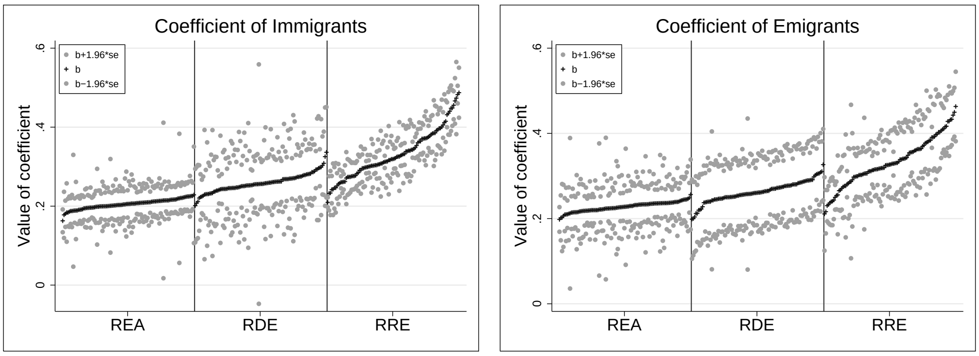
[21] Results are virtually unchanged with the alternative measure of distance, i.e. aggregate distance.

It can readily be seen from table 15 that, controlling for networks reduces the distance elasticity by about one-third, whereas the contiguity effect is three to four times smaller. The border effect is reduced even further, and disappears completely at the RE-LS scales. The MAUP distortions are subsequently far larger than those observed in table 13 and 14.

It is worth noting that the trade-creating effect of migrants is robust to the shift of zoning system, in terms of both size and shape. Migrant and business network variables are indeed summed from one scale to another, and this aggregation process increases both their mean and dispersion. Their elasticity is not very sensitive to the MAUP because the dependent variable, trade, is aggregated under the same summation process. By way of contrast, even though the trade-creating impact of business networks increases slightly with the scale of administrative units, it is no longer statistically significant for grid zoning systems.

Figure 9 displays the estimated immigrant and emigrant coefficients in the same way as in figure 8. Both groups of estimates monotonically increase with the level of aggregation.

Figure 9 – The size- and shape-dependency of the trade-creating impact of migrants



Note: (*i*) The coefficients (*b*) have to be greater (in absolute value) than 1.96 times the standard error (se) to enter into the 95% confidence interval. (*ii*) (REA): Random employment areas, (RDE): Random Départements, (RRE): Random Régions.

We therefore continue to find that size matters more than shape. Moreover, the magnitude of this distortion is definitely larger than in our previous exercises. The explanation is that gravity regressions involve variables aggregated under different processes. Since the MAUP is fundamentally linked to whether the distribution of variables is preserved, it jeopardizes gravity estimations more than wage equations. Still, MAUP distortions remain of smaller magnitude than mis-specification biases.

7. Conclusion

The overall picture is fairly clear. The use of different specifications to assess spatial concentration, agglomeration economies, and trade determinants produces substantial variation in the estimated coefficients. In most cases, theory provides a clear explanation of such variations. Although the size effect of the MAUP might still be important, especially at large scales, it is of second-order compared to specification at lower scales. Shape distortions remain of only third-order concern. On the other hand, when zoning systems are specifically designed to address local questions, as is the case for French employment areas, we definitely argue that they should be used. Those who are left with other administrative units should not worry too much however, as long as the aggregation scale is not too large. We therefore urge researchers to pay attention in priority to choosing the relevant specification for the question they want to tackle.

We also want to draw attention on the fact that the aggregation process conditions the magnitude of the MAUP distortions. If these distortions are negligible when both the dependent and explanatory variables are averaged, they are clearly more jeopardizing when the aggregation processes are not consistent on both sides of the regression, and even more that we work with large-scale spatial units. For instance, the MAUP could be of greater concern with U.S. data aggregated at the State level.

We do not of course claim that the various specifications used in this paper are actually the best. They are simply those frequently found in the economic geography literature. Many other empirical questions can be considered. We focus on three simple exercises because they are quite different in spirit, and cover a wide range of estimations. This makes us fairly confident that our conclusions are robust to other exercises, even though this remains to be shown.

Finally, the French economical and institutional design may be particularly well-designed to minimize MAUP problems. For instance, the division of France into Départements, was adopted simultaneously with the first French constitution in 1790 to replace the old “provinces”, which more or less represented dioceses. These latter exhibited significant variation in tax systems, population and land areas, and the new division aimed to create more “regular” spatial units under a common central legislation and administration. Their size was chosen so that individuals from any point in the Département could make the round trip by horse to the capital city in no more than two days, which translated into a radius of 30 to 40 km. Hence, it might well be that the French administrative zoning systems are less sensitive to the MAUP by definition. We therefore encourage researchers to replicate the exercises carried out here in the context of other countries.

References

Abowd, J. M., Creecy, R., Kramarz, F., 2002. Computing person and firm effects using linked longitudinal employer-employee data. Working paper, Cornell University.

Amrhein, C., Flowerdew, R., 1992. The effect of data aggregation on a poisson regression model of Canadian migration. Environment and Planning A 24, 1381–1391.

Amrhein, C. G., 1995. Searching for the elusive aggregation effect: Evidence from statistical simulations. Environment and Planning A 27, 105–119.

Amrhein, C. G., Reynolds, H., 1997. Using the Getis statistic to explore aggregation effects in metropolitan Toronto census data. The Canadian Geographer 41 (2), 137–149.

Arbia, G., 1989. Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems. Kluwer.

Ciccone, A., Hall, R., 1996. Productivity and the density of economic activity. American Economic Review 86 (1), 54–70.

Combes, P.-P., Duranton, G., Gobillon, L., 2008a. Spatial wage disparities: Sorting matters! Journal of Urban Economics 63 (2), 723–742.

Combes, P.-P., Lafourcade, M., Mayer, T., 2005. The trade-creating effects of business and social networks: Evidence from France. Journal of International Economics 66 (1), 1–29.

Combes, P.-P., Mayer, T., Thisse, J.-F., 2008b. Economic Geography. Princeton University Press.

Devereux, M. P., Griffith, R., Simpson, H., 2004. The geographic distribution of production activity in the U.K. Regional Science and Urban Economics 34 (5), 533–564.

Duranton, G., Overman, H. G., 2005. Testing for localisation using micro-geographic data. Review of Economic Studies 72 (4), 1077–1106.

Eeckhout, J., 2004. Gibrat's law for (all) cities. American Economic Review 94 (5), 1429–1451.

Ellison, G., Glaeser, E. L., 1997. Geographic concentration in U.S. manufacturing industries: A dartboard approach. Journal of Political Economy 105 (5), 889–927.

ESPON, 2006. The modifiable areas unit problem. Tech. rep., European Spatial Planning Observation Network.

Feenstra, 2003. Advanced International Trade: Theory and Evidence. Princeton University Press.

Fotheringham, A. S., Wong, D. W. S., 1991. The modifiable areal unit problem in multivariate statistical analysis. Environment and Planning A 23, 1025–1044.

Gabaix, X., Ioannides, Y. M., 2004. The evolution of city size distributions. In: Henderson, J. V., Thisse, J. F. (Eds.), Handbook of Regional and Urban Economics. Vol. 4: Cities and Geography. Elsevier, Ch. 53, pp. 2341–2378.

Gehlke, C., Biehl, K., 1934. Certain effects on grouping upon the size of the correlation coefficient in census tract material. Journal of the American Statistical Association 29 (185), 169–170.

Hanson, H. G., 2005. Market potential, increasing returns and geographic concentration. Journal of International Economics 67 (1), 1–24.

Harris, C., 1954. The market as a factor in the localization of industry in the United States. Annals of the Association of American Geographers 44, 315–348.

Holmes, T. J., Lee, S., forthcoming. Cities as six-by-six-mile squares: Zipf's law? In: Glaeser, E. L. (Ed.), The Economics of Agglomeration. University of Chicago Press.

Holt, D., Steel, D., Tranmer, M., Wrigley, N., 1996. Aggregation and ecological effects in geographically-based data. Geographical Analysis 28, 244–261.

McCallum, J., 1995. National borders matter: Canada-U.S. regional trade patterns. American Economic Review 85 (3), 615-623.

Maurel, F., Sédillot, B., 1999. A measure of geographic concentration in French manufacturing industries. Regional Science and Urban Economics 29 (5), 575–604.

Menon, C., 2008. The bright side of MAUP: an enquiry on the determinants of industrial agglomeration in the United States. Working Paper n. 29/2008 of Department of Economics, Ca'Foscari University of Venice.

Openshaw, S., Taylor, P., 1979. A million of so correlation coefficients: Three experiments on the modifiable areal unit problem. In: Wrigley, N. (Ed.), Statistical Applications in the Spatial Sciences. Pion London, pp. 127–144.

Rauch, J., 2001. Business and social networks in international trade. Journal of Economic Literature 39 (4), 1177–1203.

Redding, S., Venables, A. J., 2004. Economic geography and international inequality. Journal of International Economics 62 (1), 53–82.

Reynolds, H. D., 1998. The modifiable areal unit problem: Empirical analysis by statistical simulation. Ph.D. thesis, University of Toronto, Department of Geography.

Wagner, D., Head, K., Ries, J., 2002. Immigration and the trade of provinces. Scottish Journal of Political Economy 49 (5), 507–25.

Appendix: Data

Economic variables for all zoning systems are obtained by aggregating information over the 36,247 French municipalities ("communes").

First, over the 1976-1996 period, the composition in terms of establishments (employment size, and number of establishments) and workers (year and place of birth, age, gender, occupation, and wage, among others) is available at the 4-digit industrial level. The data come from the INSEE survey "*Déclaration Annuelle de Données Sociales*" (DADS), which collects matched employer-employee information in France. Our analysis builds on a panel extract covering people born in October of all even-numbered years, excluding civil servants, which is a representative $1/24^{th}$ of the French population. No survey was carried out in 1981, 1983 or 1990, producing a final sample of over 12.3 million plant-individual year observations, which are then re-aggregated by spatial unit, year (18 points), and industry (98 two-digit sectors covering both manufacturing and services).[22] As the key parameter of the sampling process is the date of birth, there is no obvious reason to believe that the sample is geographically biased.

For 1996, the above data are matched with information on the trade volumes shipped by road, both

[22] As in Abowd et al. (2002), part-timers are retained and outliers (over five standard errors above and below the mean) are dropped. The selection of industries and the removal of sampling errors at the smallest scale follows Combes et al. (2008a).

within and between municipalities, which we aggregate into different larger zoning systems. The data comes from the French Ministry of Transport, which annually surveys a stratified random sample of trucks.

Regarding social and business networks, we compute migrant stocks based on the number of natives from one area who moved to work in another area.[23] Business networks are captured via the number of financial connections between plants belonging to the same business group. For each business group, we count the number of plants located in each area. We then compute for each pair of areas the sum over all business groups of the product of the two counts. The data source here is the INSEE survey "*LIaisons FInancières*" (LIFI), which defines a business group as the set of all firms controlled either directly or indirectly (over 50%) by the same parent firm, which is itself not controlled by any other firm.[24]

[23] This figure is also calculated using the DADS survey.

[24] See Combes et al. (2005) for more details on the network variables.