Automatic delimitation of labour market areas based on multi-criteria optimisation: the case of Spain 2011

Lucas Martínez-Bernabéu

International Economics Institute, University of Alicante, Spain.

José Manuel Casado-Díaz

International Economics Institute and Department of Applied Economic Analysis, University of Alicante, Spain.

Abstract

Labour market areas (LMAs) are a type of functional region (FR) defined on commuting flows and used in many countries to serve as the territorial reference for regional studies and policy making at local levels. Existing methods rely on manual adjustments of the results to ensure high quality, making them difficult to be monitored, hard to apply to different territories, and onerous to produce in terms of required work-hours. We propose an approach to automatise all stages of the delineation procedure and improve the final results, building upon a state-of-the-art stochastic search procedure, that ensures optimal allocation of municipalities/counties to LMAs while keeping good global indicators: a pre-processing layer clusters adjoining municipalities with strong commuting flows to constrain the initial search space of the stochastic search, and a multi-criteria heuristic corrects common deficiencies that derive from global maximisation approaches or simple greedy heuristics. It produces high quality LMAs with optimal local characteristics. To demonstrate this methodology and assess the improvement achieved, we apply it to define LMAs in Spain based on the latest commuting data.

Keywords: Functional Regions, Functional Economic Areas, Commuting Zones, Labour Market Areas, Optimisation Problem.

To cite this article:

Martínez-Bernabéu, L., & Casado-Díaz, J. M. (2022). Automatic delimitation of labour market areas based on multi-criteria optimisation: The case of Spain 2011. *Environment and Planning B: Urban Analytics and City Science, 49*(2), 654–670 (First Published June 4, 2021).

https://doi.org/10.1177/23998083211021104

Supplemental material: https://journals.sagepub.com/doi/suppl/10.1177/23998083211021104

1. Introduction

1.1. Background

Labour Market Areas (LMAs), understood as the functional regions (FRs) in which supply and demand for labour meet and fix a price (Brown and Holmes, 1971; Jones, 2017), are a useful instrument to analyse the economic reality of the territory and to design, apply and monitor socio-economic policies at the regional level (Casado-Díaz and Coombes, 2011; OECD, 2002; Wicht *et al.*, 2020). Other examples of FRs, particularly within the concept of functional economic areas (FEAs) (Jones, 2017), include functional urban areas and metropolitan areas (OECD, 2012), and housing market areas (Jones *et al.*, 2012). The precision of regional analyses and the effectiveness of the respective public policies depend on how accurately the boundaries of the regions on which they are based reflect the actual spatial functional reality of the territory. Hence, assuring high standards of quality in the definition of such FRs is of great relevance.

LMAs are composed of basic spatial units (BSUs), such as municipalities, counties or census areas, which are grouped using *regionalisation* methods that process information on the commuting flows existing between them. In an ideal LMA, there is a strong interaction (*cohesion*) between the BSUs that integrate the region, and weak interaction with other regions –i.e. the region has a high degree of *autonomy*– (Goodman, 1970). These objectives are in conflict: the maximum autonomy is reached when a single region encompasses the whole territory, which implies insufficient cohesion because not every job is accessible to every worker (too large regions, excessive travel distances), and, conversely, the maximum degree of cohesion implies small regions with boundaries frequently crossed by work commuters (insufficient autonomy). This trade-off brings uncertainty to what should be considered an appropriate compromise between cohesion and autonomy. When conducting a delimitation of LMAs, most methods accept one or more parameters to control how the conflicting objectives are weighted. Most methods also require the regions to be spatially continuous. In Martínez-Bernabéu *et al.* (2020), the authors review the concept and use of LMAs and elaborate on the complexity of the problem.

1.2. Regionalisation methods

Defining optimum LMAs for a given territory (i.e. finding the division of that territory into relatively autonomous and cohesive regions) is a complex exercise in real world. Exact methods are not computationally affordable except for small enough instances. For example, Kim *et al.* (2015) use linear programming to define FRs in Seoul (South Korea), with 25 BSUs, and South Carolina (USA), with 46 BSUs. The process required minutes for the former and hours for the latter. The computational

time required grows exponentially with the number of BSUs, with tens of thousands in common study cases.

Greedy algorithms, suboptimal but capable of solving real-world problems in reasonable time, have been traditionally used in this field (Masser and Scheurwater, 1980). Most of these methods are either pure hierarchical clustering methods such as Intramax (Masser and Brown, 1975) or pseudo-hierarchical such as the well-known *Travel-to-Work Areas* method (last revision in Coombes and Bond, 2008). Their aggregative procedure is guided by an interaction index that measures the strength of commuting between two areas considering the number of commuters shared relative to the size of the two areas.

Recently, approximate methods based on optimisation of a global objective function have shown to find better solutions. Compared with greedy methods, approximate procedures require more computational effort, but in exchange they are able to find results closer to optimum (Coombes *et al.*, 2012; Fowler and Jensen, 2020; Martínez-Bernabéu *et al.*, 2020). These methods are a good compromise between the fast but less effective greedy methods and the optimum but computationally unaffordable exact methods. Two objective functions are predominant in this field: modularity quality (Mu and Yeh, 2020; Shen and Batty, 2019), and a form of global cohesion that measure average BSU-LMA interaction (Alonso *et al.*, 2015; Flórez-Revuelta *et al.*, 2008; Martínez-Bernabéu *et al.*, 2012). Modularity can fail to identify very large or very small communities, even when a resolution parameter is arbitrarily modified to accommodate the scale of division to the practitioners expectations (Lambiotte, 2010; Lancichinetti and Fortunato, 2011)). Average BSU-LMA interaction is biased towards high cohesion regionalisations with many small LMAs (Casado-Díaz *et al.*, 2017a; Martínez-Bernabéu *et al.*, 2020), but a constraint on minimum LMA autonomy and/or size makes it possible to produce results of any level of detail as desired, based on readily understood indicators (autonomy, size, area) instead of an obscure resolution parameter that doesn't relate to the regions characteristics sought in these exercises.

1.3. Deficiencies in the results of functional regionalisation methods

The revision of existing methods, especially those used by public administrations, shows that their outcomes are very frequently subject to expert examination and rounds of consultations with local authorities, and the procedures include final steps of manual adjustments, usually undocumented (Casado-Díaz and Coombes, 2011). See Franconi *et al.* (2017) for an example of such correction procedures. The most common corrections are reallocation of BSUs to different LMAs to improve

their local characteristics (increasing its autonomy within its LMA and/or the interaction with the other BSUs in it), division of too big LMAs, and merger of too small LMAs functionally dependent on each other. These changes are always based on expert knowledge that supersedes the regionalisation algorithm criteria, and onerous when there are hundreds or thousands of BSUs. A complete system to automatise such procedures in a consistent and replicable manner would imply an increase in the quality of the results and savings in time and money.

The issues that require a final phase of manual correction can be grouped according to this typology:

- Cohesion-misallocated BSUs: those that adjoins a LMA to which they could be reallocated to achieve stronger interaction.
- Autonomy-misallocated BSUs: those that adjoins a LMA to which they could be reallocated to achive higher autonomy (optionally, "without causing significant loss in cohesion").
- Convenience LMAs: regions that barely meet the constraints of autonomy and size to be a valid LMA and that are composed by one or more misallocated BSUs whose removal would render the LMA invalid.

1.4. Automatising the regionalisation process

The aim of this work is to present an automatic method that tackles the whole problem of LMA delimitation, from data pre-processing to the final local corrections, starting from an existing regionalisation method. We develop a multi-step procedure around the Grouping Evolutionary Algorithm (GEA) proposed by Casado-Díaz *et al.* (2017b), a refinement of the method by Martínez-Bernabéu *et al.* (2012). We demonstrate the suitability of our proposal by applying it to the commuting data from Spanish 2011 Census to define LMAs, and evaluate the results using a set of quantitative indicators extracted from relevant literature (Martínez-Bernabéu *et al*., 2020).

In the rest of the paper, Section 2 describes the data, Section 3 formally describes the problem at hand and the methodology we propose, Section 4 discusses the results with and without the proposed improvements, and Section 5 concludes.

2. Data

The main data required to solve this problem is a non-symmetrical matrix T of travel-to-work (commuting) flows between the set S of n BSUs, where T_{ij} is the number of residents of BSU s_i that work in BSU s_j. These values derive from the questions on place of residence and place of work from the latest Spanish Census (INE, 2011). The smallest available BSU is the municipality (numbering

8,116). To be able to apply size restrictions on population size (inhabitants), $P=\{p_1, p_2, \dots p_n\}$ was derived from the same source, where p_i is the population of BSU i. Additionally, the algorithm requires the data on geographical adjacency between each pair of BSUs if the contiguity restriction is to be applied. We created this database, a binary symmetrical matrix, using geographic information systems in R (R Core Team, 2013). Figure S1 in the Supplementary Material exemplifies a commuting and an adjacency matrix from a synthetic territory.

It should be noted that previous censuses performed an exhaustive coverage of all Spanish population. That was not the case of the 2011 wave when, in line with other countries, the INE adopted a sample-based approach in which less than 10% of the population answered the census questionnaire. Furthermore, the estimated BSU-to-BSU commuting flows were rounded to multiples of 5 due to significance and confidentiality concerns. These two facts challenged the statistic quality of the data compared with previous exercises, with a stronger effect for less populated BSUs. In particular, many small BSUs (more than 10%) did not register travel-to-work flows to any of their neighbouring BSUs, while in 2001 this happened in less than 1% of cases. This exemplifies a development in the data quality that deserves attention given its progressive generalisation and potential implications (United Nations Statistics Division, 2019).

3. Methodology

In this section we formally describe (a) the regionalisation problem as an optimisation problem, (b) the GEA method used as the regionalisation algorithm around which we build our multi-step procedure, and (c) the algorithms for the two extra steps we propose in this paper. Table S1 in the Supplementary Material lists all acronyms and symbols used in this section, including the parameters of each algorithm.

3.1 Regionalisation as an optimisation problem

The aim is finding the partition R of a set of BSUs $S = \{s_1, s_2, \dots s_n\}$ into geographically contiguous, non-overlapping and sufficiently autonomous and populated regions (LMAs) that maximise *cohesion* measured as the average interaction between each BSU and the rest of its LMA (Flórez-Revuelta *et al.*, 2008). Formally:

Maximise

$$f(R) = \sum_{M \in R} \left(\sum_{s \in M} \text{interaction}(s, M \setminus s) \right) \quad (1)$$

subject to

$$\cup_{i=1..m} M_i = S \tag{2a}$$

$$M_i \cap M_j = 0 \forall M_i \neq M_j \in R \tag{2b}$$

$$\text{contiguous}(M_i) = 1 \forall M \in R \tag{2c}$$

$$\sum_{k \in M} p_k \geq P_{\min} \forall M \in R \tag{3a}$$

$$\text{autonomy}(M) \geq A_{\min} \forall M \in R \tag{3b}$$

$$\text{autonomy}(M) \geq A_{\text{tar}} + \frac{A_{\text{tar}} - A_{\min}}{P_{\min} - P_{\text{tar}}} \left(\sum_{k \in M} p_k - P_{\min} \right) \forall M \in R \tag{3c}$$

where M is each of the m regions of partition R; contiguous(M) equals 1 if the BSUs of M form a single geographically contiguous region and 0 otherwise; $P_{\min} < P_{\text{tar}}$ and $A_{\min} < A_{\text{tar}}$ are the parameters of, respectively, minimum population, target population, minimum autonomy and target autonomy of a LMA to be considered valid.

Eqs. (2a-c) ensure that a regionalisation is a partition of the whole territory (Eq. 2a) into non-overlapping (Eq. 2b), contiguous (Eq. 2c) regions. Eqs. (3a-c) conform the validity criteria: each region must reach the required minimum levels of $P_{\min}$ inhabitants (Eq. 3a) and A_{tar} autonomy (Eq. 3b), linearly relaxed to $A_{\min}$ for regions that reach P_{tar} or more inhabitants (Eq. 3c). Figure S2 in the Supplementary Material illustrates the validity criteria.

The autonomy of a region is measured as the number of residents of that region that hold a job also in that region divided by the maximum between total occupied residents and total jobs of that region:

$$\text{autonomy}(M_x) = \frac{\sum_{i \in M} \sum_{j \in M} T_{\text{ij}}}{\max\left(\sum_{i \in M} \sum_{k \in S} T_{\text{ik}}, \sum_{j \in M} \sum_{k \in S} T_{\text{jk}} \right)} \tag{4}$$

The interaction between two regions is measured using the following index (a modification of the one used in the *Travel-to-Work Areas* method, proposed by Casado-Díaz *et al.*, 2017b, to reduce the presence of cohesion-misallocated BSUs):

$$\text{interaction}(M_a, M_b) = \frac{1}{2}\sqrt{\frac{\left(\sum_{i\in M_a}\sum_{j\in M_b} T_{\text{ij}}\right)^2}{\sum_{i\in M_a}\sum_{k\in S} T_{\text{ik}} \cdot \sum_{j\in M_b}\sum_{k\in S} T_{\text{kj}}} + \frac{\left(\sum_{i\in M_a}\sum_{j\in M_b} T_{\text{ji}}\right)^2}{\sum_{j\in M_b}\sum_{k\in S} T_{\text{jk}} \cdot \sum_{i\in M_a}\sum_{x\in S} T_{\text{ki}}}} \quad (5)$$

In the rest of the paper, we use Eq. (5) as the index to determine when a BSU is cohesion-misallocated as well as in the objective function. To assess autonomy-misallocation, we calculate the minimum of supply-side and demand-side autonomy that the BSU would reach on each LMA, and call it *within-region autonomy*:

$$w(s_i, \text{M}_x) = \text{minimum}\left(\frac{T_{ii} + \sum_{j\in M_x} T_{\text{ij}}}{\sum_{k\in S} T_{\text{ik}}}, \frac{T_{ii} + \sum_{j\in M_x} T_{\text{ji}}}{\sum_{k\in S} T_{\text{ki}}}\right) \quad (6)$$

3.2 Multi-step procedure built around an optimisation algorithm

GEA (Casado-Díaz *et al.*, 2017b) is a state-of-the-art regionalisation method that has proved to be superior to other pre-existing, relevant formal procedures (see Martínez-Bernabéu *et al.*, 2012, for the Spanish case; Coombes *et al.*, 2012, for a comparison of GEA, the Swedish and the Travel-to-Work Areas methods applied to Spain, Sweden and UK; and Casado-Díaz *et al.*, 2017b, for the Chilean case). In short, this method uses parallel evolutionary computation to perform a stochastic search of the solution space. Using several sets (*cells*) of possible regionalisations, arranged in a toroidal grid, each cell of the grid is populated at start with different regionalisations produced with a greedy stochastic hierarchical aggregative procedure. GEA then applies group-based mutation and crossover operators to produce a new regionalisation within each grid cell, ranks it using its global objective function (Eq. 1), and removes the one of the less successful regionalisations from each cell. At a lower frequency, a regionalisation from one cell is crossed with a regionalisation from a cell adjacent in the grid. The process continues until no improvement of the best result in the whole grid is found for a certain number of iterations or a maximum computation time is reached. Figures S3 and S4 in the Supplementary Material illustrates GEA data structures and algorithm.

As happens with any other functional regionalisation method, GEA results can include cases of the deficiencies commented in Section 1.3. Casado-Díaz *et al.* (2017b) attributed the presence of cohesion-misallocated BSUs in the results of a method that precisely optimises (global) cohesion to the nature of the grouping optimisation process: reallocating any of these BSUs to its optimal LMA would cause a slight drop in the interaction index between other BSUs and their respective LMAs, an effect that once aggregated in the global objective function offsets the improvement in interaction from the reallocated

BSU, and so the method keeps such BSUs misallocated. The existence of autonomy-misallocated BSUs is intrinsic to this problem due to the conflict between autonomy and cohesion: since this method maximises cohesion, some or many BSU could show suboptimal autonomy and optimal cohesion. Fixing one would break the other. Autonomy-misallocated BSUs are bad only when the potential gain in autonomy is great and the potential loss in cohesion is small. Convenience LMAs seem to arise from the previous two issues.

In order to address these limitations, we propose to wrap GEA into a pre-processing layer and a post-processing layer, to form a three-step method that deals with the issues with different approaches (Figures S5 and S6 in the Supplementary Material illustrates the original and the proposed methodologies):

1. A hierarchical aggregation of BSUs, driven by percentage of commuters, merges pairs of areas that share great amounts of their workers. This introduces a form of BSU-level, soft constraint that minimises the appearance of autonomy-misallocated BSUs as well as convenience LMAs in GEA results.
2. GEA is applied to the pre-processed matrix of flows. Constrained by the aggregated network, the search focuses on producing more suitable regionalisations. Reducing the search space dimension accelerates the computation as well.
3. A greedy local optimisation algorithm applied to the GEA results performs simple reassignments of misallocated BSUs in order of potential interaction gain, subject to constraints of cohesion and autonomy at both global and BSU level. This stage corrects the remaining issues at local level and allows reconsideration of the mergers done in first step.

3.2.1. Pre-processing layer

In order to reduce the occurrence of autonomy-misallocated BSUs, the pre-processing layer follows a simple principle: merge the pair of adjacent BSUs that maximise Eq. (6) if it exceeds a given threshold $1 - A_{min}$ (e.g. for A_{min}=66.7% we get 33.3%), and in that case recalculate commuting flows to and from the new combined area and repeat.

The motivation is as follows: If the share of commuters of a BSU that work in a given region is d, the maximum within-region autonomy for that BSU if it is not allocated to that region will be $1 – d$. By setting the minimum threshold at $1 - A_{min}$, it is ensured that every BSU with such a strong dependence to another BSU (or subset of BSUs) will be allocated to the same LMA.

The algorithm can be briefly described as follows:

Step 1: find the pair of adjacent BSUs a and b that maximises Eq. (6).

Step 2: if $w(a, b) > 1 - A_{min}$, merge a and b and update the adjacency and commuting matrices accordingly.

Step 3: if there was any merge in last iteration, go back to step 1.

The complete pseudo-code of the pre-processing algorithm is included in the Supplementary Material, Method 1, as well as a diagram, Figure S7. Figure S1 includes an example of how the method transforms the data.

3.2.3. Local optimisation layer

The final step of local optimisation aims at solving all the issues highlighted in Section 1.3 that could remain in the regionalisation produced by the GEA method. For that, it overrides the global objective function and focus on reaching a proper balance between global quality and local adequateness. The formulation of the rules that drive this process is relatively complex to address more complicated and ambiguous trade-offs between weighted interaction and absolute commuting dependence of BSUs and LMAs. The process is as follows:

Phase 1: Ranking of BSUs.

All BSUs are ranked in descending order of *potential cohesion increase* (PCI), i.e. the increase of Eq. (5) if BSU was reallocated to the best adjacent LMA. All BSUs are marked as *unchecked*. BSUs with positive PCI are cohesion-misallocated.

Phase 2: Main loop.

Consider the unchecked BSU with maximum positive PCI. If the reallocation of the current BSU would break validity of the receiving LMA, the BSU is marked as checked and next one is considered. Otherwise, the reallocation is evaluated if any of the following conditions is met:

- the current (donor) LMA would be valid after reallocation or is already invalid, or
- the BSU is not the only one in the current LMA and its interaction with that LMA is minimal (<0.0001), or
- reallocating the BSU increases its within-region autonomy from below to above A_{min}, or by more than $1-A_{tar}$ (i.e. >25% with the validity parameters used in this work).

In other words, the reallocation is *tentatively* performed if the BSU has no significant interaction

with the rest of its LMA and was allocated to it by convenience (excluding LMAs formed solely by the BSU considered, which are always kept if valid), in which case the reallocation is allowed to break validity of the donor LMA, or if the increase in within-region autonomy for the BSU is significant (enough to get it above A_{min} or as big as to prevent reaching A_{tar}).

This process affects the PCI and autonomy of other BSUs. Some of those might become misallocated, so a process begins to check whether the new result is preferable (except when the donor LMA was invalid before reallocation, in that case the reallocation is accepted directly without further checks): for every BSU affected by the last reallocation (any BSU in the affected LMAs or sharing commuters with them), calculate worst (maximum) and best (minimum) PCI change, and final maximum PCI among the affected BSUs. The reallocation is definitely accepted if:

- the new maximum PCI is smaller than before reallocation (none of the affected BSUs is now in a worse state than the initially reallocated BSU was) and one of the following two conditions are met:
    - the global fitness, Eq. (1), is greater now and the best PCI change is greater in absolute value than the worst change, or
    - the within-region autonomy, Eq. (6), of the reallocated BSU has increased in more than 1-A_{tar}.

If this process breaks the validity of the original LMA, it is considered a convenience LMA, divided into its internally contiguous groups of constituent BSUs, and the BSUs of groups that cannot constitute a valid LMA are reallocated to their best adjacent LMA, in descending order of PCI, until all the remaining group are again valid LMAs or disappear. All the BSUs affected by the reallocation are marked as *unchecked*, to be considered again for reallocation if they are cohesion-misallocated. Finally, the BSU considered in this iteration (reallocated or not) is marked as *checked* and the loop begins again, until all BSUs with positive PCI are *checked*.

The complete pseudo-code of the optimisation algorithm is included in the Supplementary Material, Method 2, as well as a diagram, Figure S8.

4. Results and discussion

In this section we analyse the results of applying the proposed methodology to a real case study – the definition of LMAs for Spain using the latest commuting data, 2011 – and compare them with the results obtained by GEA without the proposed extra layers. All the algorithms were written in C and

run in an Intel i5-2500 CPU at 3.3 GHz with 16 GB of RAM over Ubuntu 18.04.

4.1. Pre-processing layer results

In this section we compare the GEA results of regionalising Spain using the original flow matrix and the compacted matrix produced by the pre-processing algorithm. We produced two sets of 20 independent GEA results, each set fed with each matrix.

The preprocessing of the matrix flow took a few seconds, with estimated complexity O(n log n). The initial merger of BSUs, using the parameter threshold of 33.33% minimum dependency (this threshold value is taken from the LMA validity criteria, it is not a parameter to be set independently), reduced the number of BSUs from 8,116 to 4,975 (4,642 were composed of a single BSU). See Figure S9 in the Supplementary Materials. The mergers of BSUs concentrate on areas around large employment centres and sporadically in some urban regions where BSUs don't have a clear dominant centre or whose main attractor is not contiguous so no merger is possible. The biggest change is the combined region formed around Madrid (consisting of 262 municipalities), whose size is comparable to the Madrid LMA with the original methodology with 2001 data (Martínez-Bernabéu *et al.* 2012), with a noticeable growth southwards.

The restriction parameters on LMA's minimum autonomy and population were set to the values used in the 2001's exercise ($A_{\min}$ = 66.7%; A_{tar} = 75%; P_{min} = 20,000; and P_{tar} = 100,000), that are also a sort of international standard (Coombes *et al*. 2012, Franconi *et al*. 2017).

We tested different values for the evolutionary parameters of GEA that influence its search efficiency: rows and columns of the grid of cells, number of regionalisations per cell, frequency of migration between cells, and the stopping condition. Few grid cells of few individuals in a non-square grid, with infrequent migration between cells and long execution times were the more successful. We used a grid of 5×2 cells, 3 regionalisations per cell, probability of migration between cells per generation 0.001, and termination after 2 hours of execution. When applying this methodology to other datasets, the rule of thumb is to start with small values (faster results) and try bigger values until best results do not vary.

Figure S10 of the Supplementary Material shows the whole maps of the best result from each set. The compacted matrix, with 38.7% fewer nodes (BSUs), allows on average 2.4 GEA iterations for the same time. This implies a significant speed-up for large case studies, such as those using smaller resolution of the commuting matrix (e.g. census districts or mobile phone cell grids as BSUs), or analysing several countries together, e.g. delineating transnational LMAs in the European Union (see Coombes *et al.*, 2012).

Table 1 shows key quality statistics (Martínez-Bernabéu *et al.*, 2020) calculated over each set of regionalisations. The first two rows correspond to the best outcome of each set, and the next two lines to the average of the 20 outcomes in each set. The last two rows correspond to the optimised regionalisations (see Section 4.2). The first statistic (#LMA) is the number of LMAs identified. Provided the same level of autonomy is reached, a larger number of LMAs implies better levels of cohesion and detail. The second statistic (A_{gbl}) is the global autonomy, the ratio of residents that work in the same LMA they reside in; higher is better, but in conflict with cohesion. The number of LMAs resulting from all these exercises is quite similar, in the range [290-306]. The average values in the outcomes based on the compacted matrix show slightly fewer LMAs (although in its best result the value reached in this indicator is among the highest). However, the pre-processing layer results in a considerable improvement of the global autonomy indicator, even for outcomes with a comparable number of LMAs. This increase of autonomy is also noticeable for individual LMAs. To analyse the less autonomous LMAs without being subject to the influence of outliers (as happens with the minimum), the third statistic displayed is the first decile of LMA autonomy. The better values for compacted results imply that the increase in global autonomy has not been achieved only through the expansion of the largest LMAs at the expense of smaller ones: the autonomy levels of almost all LMAs are either comparable to those in the original case or have improved.

Table 1: Regionalisation quality statistics

REGIONALISATION	#LMAs	A_{gbl}	A_{LMA} 1st decile	#BSU aut-mis	A_{BSU} 1st decile	#A_{BSU}<.5	Cohesion	#BSU coh-mis	Area max.
Original (best)	298	89.35%	73.76%	666	57.14%	461	.036095	267	9612
Compacted (best)	302	91.79%	74.33%	565	59.26%	382	.036069	359	12168
Original (mean)	301.7	89.95%	74.06%	700	57.47%	446	.035905	297	9684
Compacted (mean)	294.3	92.01%	74.59%	561	60.46%	345	.035645	342	12357
Original (best) opt.	286	89.49%	73.64%	603	57.85%	436	.035804	162	9419
Compacted (best) opt.	290	91.79%	74.30%	563	58.56%	408	.036038	173	10302

The autonomy levels of each BSU within its LMA are also better for compacted results, as indicated by the next three statistics in Table 1: the number of autonomy-misallocated BSUs (#BSU aut-mis) is lower; the first decile of BSU autonomy (A_{BSU} 1st decile) is greater, and the number of BSUs with less than 50% autonomy in its LMA (#A_{BSU}<0.5) is lower.

The sixth indicator in Table 1, average BSU-LMA interaction (Cohesion), shows similar values for both sets, slightly better for the original results, as expected since solving autonomy-misallocated BSUs takes a toll on average cohesion. The only indicator for which the original results are clearly superior is the number of cohesion-misallocated BSUs (seventh indicator, #BSU coh-mis). This seems unavoidable, since autonomy-misallocated BSUs in the original results are cohesion-misallocated in the compacted results, and *vice versa*. However, as indicated by the similar values of cohesion, the trade-off is better for the compacted results: the increase in autonomy of cohesion-misallocated BSUs compensates for the reduction in cohesion.

Overall, compacted results show a considerable improvement in autonomy at all resolution levels, for both LMAs and BSUs. BSUs are more dependent on their LMAs and less on other LMAs. Moreover, such improvement is reached while maintaining a similar number of LMAs and average cohesion. This is actually an achievement since better local characteristics of the BSUs were expected to cost a worsening of global interaction and number of LMAs.

The last statistic in Table 1, the area of the largest LMA (Area max.), reflects the most visual effect of using the compacted matrix: the southward expansion of Madrid LMA (see Figure S11 in the Supplementary Material). This could be an undesired result if it wasn't justified by the change in commuting patterns from 2001 to 2011, but the proposed layer of final optimisation is able to correct any allocation forced by the pre-processing layer, as shown in subsection 4.2.

A more fine-grained examination of the results, at individual LMA and BSU level, also favours the compacted results: The "worst" allocations of BSUs in the compacted results correspond to municipalities in the lowest ranks of population for which there is no alternative LMA that could improve their local characteristics (they show functional dependence to far, non-adjacent LMAs), while the original results include some grievous misallocations with immediate solutions. The same applies to convenience LMAs, much less frequent in the compacted results, precisely because the merged BSUs prevent those borderline solutions. Figure 1 (a) illustrates this: In the original results, the municipality Albocàsser is included in a LMA whose population is just above the threshold (20,429) and has a relatively tortuous shape. Albocàsser has low autonomy within that LMA (45.78%), which would be much greater if reallocated to the neighbouring LMA of Castelló de la Plana (86.12%). The interaction index also favours Castelló de la Plana (0.02539 vs. 0.02059). However, the reallocation breaks validity of Alcalà de Xivert LMA, and the reallocation of the remaining BSUs causes a net loss of the objective function even if less BSUs are misallocated. This is an example of a convenience LMA formed to get a marginal gain in the objective function in expense of worse local characteristics. Such

objective function's artefact is avoided by the pre-processing layer because Albocàsser is merged to Castelló de la Plana.

Another improvement of the compacted results is that certain urban regions are ensured to form a single LMA, comparable to FRs defined for other purposes such as functional urban areas (OECD, 2012). The best example is the case of Barcelona (Figure 1b), a metropolitan area that articulates a single large LMA in the compact results but is divided into two LMAs in the original results. Such single LMA encloses most commuting interaction between its constituent BSUs. A subdivision of a large FR could be useful for specific purposes[1], provided the strong interaction between both regions is appropriately acknowledged and taken into account, but for certain analysis it is preferable to use the single region framework.

1 In the provided research data, we report the results of replicating the study by Martínez-Bernabéu and Casado-Díaz (2016) to subdivide large enough LMAs into *microclusters*, to increase territorial detail in microdata files for econometric analysis.

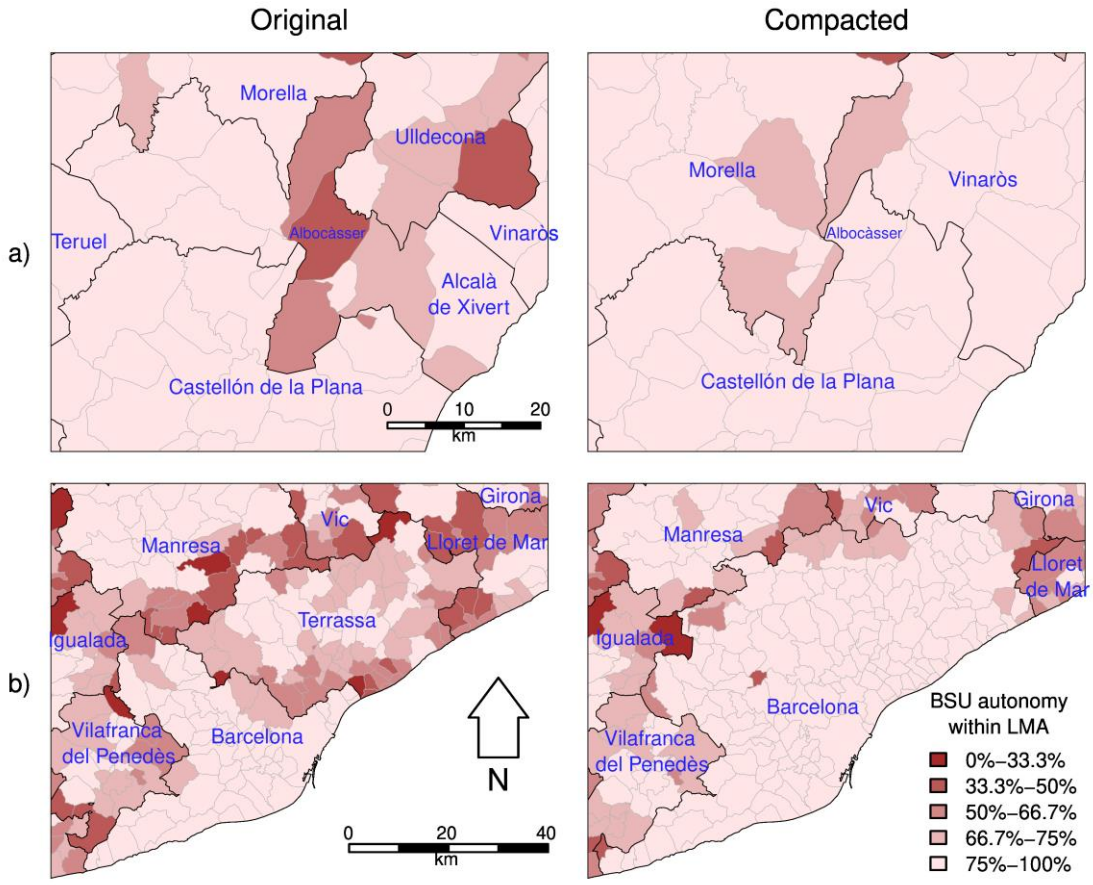

Figure 1: (a) Convenience LMA of Alcalà de Xivert in the original results does not appear in the compacted results; (b) The metropolitan area of Barcelona is divided in the original results and kept together in the compacted results.

4.2. Local optimisation results

We apply the automatic optimisation process to the best result of each of the sets of LMAs obtained in the previous section, requiring less than a minute of computations and an estimated time complexity of O(n log n). Again, there is no need to set any parameter, which are determined by the LMA validity criteria. The inclusion of the results for the original matrix allows us to compare the effects on results with different characteristics. For the compact case, this process caused the reallocation of 544 BSUs: 216 are corrections of misallocations and 318 due to the dismembering of 12 LMAs that became invalid in terms of size or autonomy after the reallocation of certain BSUs.

Las two rows of Table 1 show the relevant statistics of the optimised results. The optimisation layer caused a 4% reduction in the number of LMAs for both results. Those 12 LMAs lost some of its BSUs during the optimisation process and became invalid LMAs (the increase observed in local indicators for the reallocated BSU was sufficiently large to justify the invalidation of its original LMA according to the rules described in Section 3.2). The remaining BSUs of such invalid LMAs were reallocated to

neighbouring LMAs. Most of these BSUs had the maximum interaction with the dismembered LMA, their reallocation invariably causes a loss of cohesion due to the dilution of their commuting flows within a bigger population. This happens even when the “new” LMA absorbs all the remaining BSUs in block (and thus all its internal interaction). That is precisely what made GEA produce such results in first place: although it assigned some BSUs to a “wrong” LMA (wrong in terms of such specific BSUs’ local indicators), its outcome maximised the objective function.

The local optimisation had different effects on each of the two results. The optimised original result shows small gains in the autonomy statistics, an improvement in the number of cohesion-misallocated BSUs (-39.3%), and a negligible loss of cohesion (-0.000291, -0.8%). In turn, the compacted optimised result shows stable global autonomy and cohesion, a noticeable reduction in the number of cohesion-misallocated BSUs (-48.2%), and small losses in BSU autonomy in the lower population range. This disparity makes sense, because the original results had more room for autonomy improvement due to the objective function maximising cohesion, while in the compacted results the merged BSUs during the pre-processing ensured fewer autonomy-misallocated BSUs and, since the local optimisation uses the non-compacted matrix, those BSUs can be reallocated when the gains justify it. This happens especially in the southern parts of the Madrid LMA, where many BSUs with strong dependence towards the Madrid metropolitan area have stronger interaction to the much smaller LMAs southwards. Overall, while the number of autonomy-misallocated BSUs remains almost unchanged, there is a great reduction in the number of cohesion-misallocated BSUs (-48.2%). The net loss in cohesion, much smaller than for the non-compacted results (-0.000031, -0.09%), is explained by the trade-off between the interaction gained from the optimisation of BSU allocation and the interaction lost due to the 12 LMAs that were eliminated during the process.

The most visible effect of the optimisation in the compacted results is the great reduction of Madrid LMA’s area, which becomes comparable to the size and shape of this LMA in the original results (See Figure S12 of the Supplementary Material). This dispels doubts about the pre-processing layer introducing a bias towards larger areas.

To complete this analysis, the LMAs that were eliminated by the optimisation process were reviewed individually to assess whether such changes are actually preferred. A detailed assessment of the interactions and autonomies of the BSUs affected by the LMA eliminations confirmed that the optimised regionalisation with fewer LMAs is actually preferable.

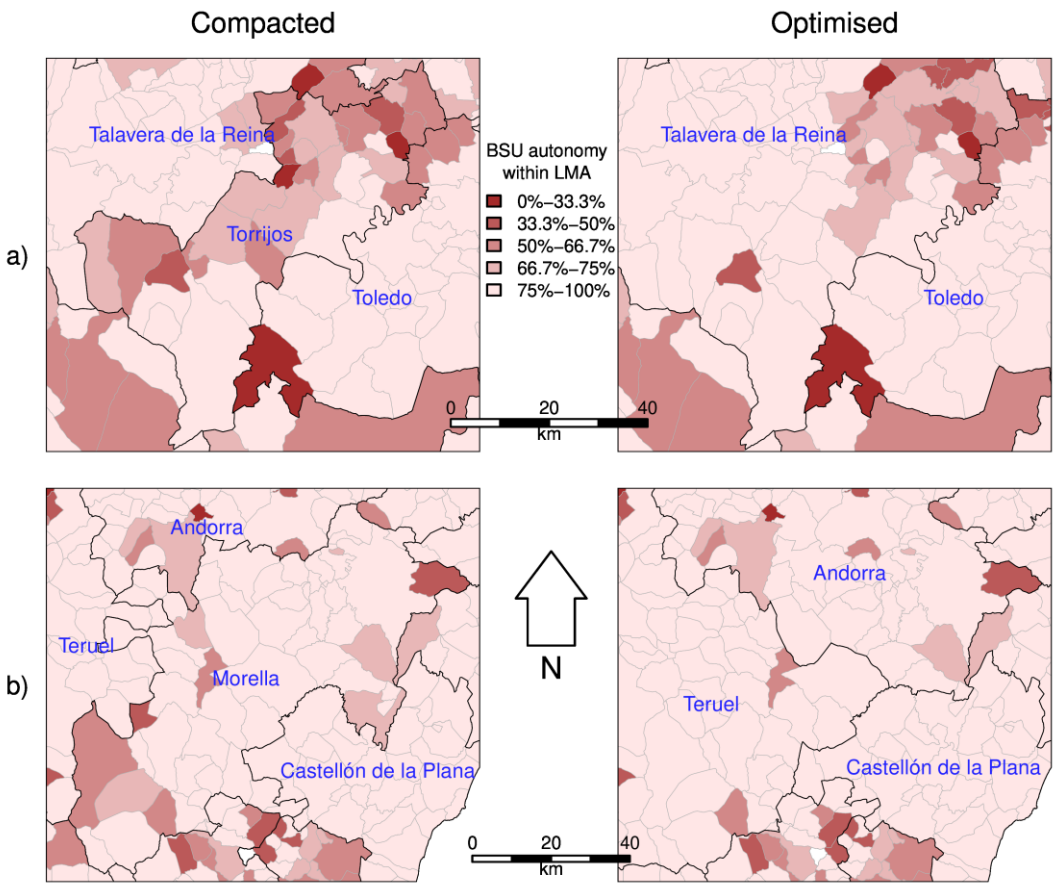

Figure 2: The optimisation process eliminates the convenience LMAs of Torrijos (a) and Morella (b) and reallocates their BSUs to the best adjacent LMAs.

Figure 2 (a) depicts the case of Torrijos LMA, the largest eliminated LMA (57,570 inhabitants) that was completely absorbed by Talavera de la Reina LMA in the optimisation process. Torrijos LMA had several misallocated BSUs. After the reallocation of such BSUs to their optimal LMA (Talavera de la Reina), Torrijos LMA no longer satisfied the minimum autonomy. Arguably, it is a convenience LMA formed by two relatively independent clusters of BSUs each of which failed to meet the minimum size and autonomy requisites to qualify as a valid LMA. The final LMA produced by the optimisation, a combination of both LMAs (Torrijos and Talavera de la Reina), has much better autonomy indicators and still qualifies as a sufficiently integrated and cohesive region.

Figure 2 (b) shows the case of Morella LMA, the most representative example of the LMAs deleted in the optimisation process (9 out of 12 cases): a small LMA that marginally meets the minimum size requisite, constituted by several sub-groups of internally cohesive BSUs that have low levels of interaction among them, which has at least one heavily misallocated BSU whose reassignment invalidates the minimum size requirement. In the case of Morella, the original LMA could be divided into two cohesive areas, north-east and south-west, with the latter showing a great dependence from a

neighbouring LMA. The optimisation performs a proper redistribution of the BSUs merging each part to a different LMA.

5. Conclusions

Providing statistics based on the appropriate FRs is important for academic research and policy making. In the case of labour force analysis, improving existing methods with new techniques is relevant to obtain LMAs' delineations that better reflect their ideal form, in a context where the relevant stakeholders face the need of dealing with increasingly greater territories, varying spatial resolutions and unreliable data.

The definition of this type of FR is tackled through the application of formal procedures that, for a given territory, group BSUs (such as municipalities or counties) into LMAs (i.e. produces a regionalisation), in such a way that each LMA is characterised by a high degree of autonomy in terms of travel-to-work (commuting) flows and the BSUs constituting it are characterised by having high levels of connection in terms of the same flows (high inner-interaction). This paper presents a three-step methodology that provides results of high quality at the aggregate, global level (number, autonomy and cohesion of identified regions) as well as at the local level (autonomy and interaction of BSUs within LMAs), greatly improving the results of previous proposals. This is achieved through the addition of an initial pre-processing layer and a final greedy, multi-criteria local-optimisation layer to a state-of-the-art grouping evolutionary algorithm that maximise average inner-interaction.

The comparison of the results obtained with and without the proposed extra steps, using the latest commuting data available for Spain, evidences the improvements achieved: The approach produces outcomes closer to the ideal of FRs in several aspects, especially in terms of BSU allocation (better trade-off between autonomy and interaction to its LMA), and the coherence of LMAs' boundaries. For roughly similar numbers of LMAs, the outcome is characterised by greater shares of commuters between BSUs and their own LMA, fewer BSUs with significant commuting flows to external LMAs, significantly less autonomy-misallocated and cohesion-misallocated BSUs, and fewer convenience LMAs. Overall, LMA boundaries are more coherent with the expert expectations. Moreover, the complete regionalisation process is considerably faster and easier: The computations performed to produce the raw results require less time and the eventual necessity of a final step of manual adjustments requires significantly less effort. Therefore, less material and human resources are required to produce good regionalisations for greater regions in less time. All in all, the proposed methodology is closer to the ideal of an unsupervised automatic system for functional regionalisation.

To conclude, a lesson from this study is that global objective functions are not perfect and inevitably show certain undesired biases that have no obvious solution, if at all. The present proposal deals with this by adding an implicit restriction on BSU allocation and by allowing a final optimisation stage that uses composite criteria based on individual BSU characteristics to complement the global objective function. This approach might be the only one to solve or alleviate the observed biases, but more research on such functions is required. The main limitation of this methodology is that it has not been tested on different territories. Although the Spanish case is quite heterogeneous in the characteristics of the BSUs (size, autonomy, commuting patterns) and can be considered as a relatively good test-bed, we cannot claim the methodology's suitability to all kind of LMA regionalisations. In our research agenda we include the integration of the global and local optimisation approaches in a single-step method, and the performing of a systematic comparison of alternative objective functions and methodologies based on multiple national case studies. To do so, we are gathering commuting data for a set of representative case studies and comparable implementations of regionalisation algorithms, to build up a comprehensive test-bed for existing and future methodologies.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Spanish Ministry of Science, Innovation and Universities / Agencia Estatal de Investigación (AEI), and the EU ERDF, grant number CSO2017-86474-R. The funding source had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

References

Alonso MP, Beamonte A, Gargallo P and Salvador M (2015) Local labour markets delineation: an approach based on evolutionary algorithms and classification methods. *Journal of Applied Statistics*, *42*(5), 1043-1063.

Brown LA and Holmes J (1971) The delimitation of functional regions, nodal regions and hierarchies by functional distance approaches. *Journal of Regional Science*, *11*(1), 57-72.

Casado-Díaz JM and Coombes M (2011) The delineation of 21st century local labour market areas: a critical review and a research agenda. *Boletín de la Asociación de Geógrafos españoles*, (57), 7-32.

Casado-Díaz JM, Martínez-Bernabéu L and Flórez-Revuelta F (2017a) Automatic parameter tuning for functional regionalization methods. *Papers in Regional Science*, *96*(4), 859-879.

Casado-Díaz JM, Martínez-Bernabéu L and Rowe F (2017b) An evolutionary approach to the delimitation of labour market areas: an empirical application for Chile. *Spatial Economic Analysis*, *12*(4), 379-403.

Coombes M and Bond S (2008) Travel-to-Work Areas: the 2007 review. *Office for National Statistics, London*.

Coombes M, Casado-Díaz JM, Martínez-Bernabéu L and Carausu F (2012) *Study on comparable labour market areas–Final Research Report*. Eurostat - Framework contract 6001. 2008.001–2009.065, specific contract 50405.2010.004–2011.325. Available at: https://ec.europa.eu/eurostat/cros/content/study-comparable-labour-market-areas_en

Flórez-Revuelta F, Casado-Díaz JM and Martínez-Bernabéu L (2008) An evolutionary approach to the delineation of functional areas based on travel-to-work flows. *International Journal of Automation and Computing*, 05(1), 10–21.

Fowler CS and Jensen L (2020) Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environment and Planning A: Economy and Space*, 0308518X20906154.

Franconi L, Ichim D, D'Alò M and Cruciani S (2017) *Guidelines for Labour Market Area delineation process: from definition to dissemination*. Istat, Italian National Institute of Statistics, Rome, Italy. Available at: https://ec.europa.eu/eurostat/cros/system/files/guidelines_for_lmas_production08082017_rev300817.pdf.

Goodman JF (1970) The definition and analysis of local labour markets: some empirical problems. *British journal of industrial relations*, *8*(2), 179-196.

INE (2011) *Censo de Población y Viviendas 2011*. Instituto Nacional de Estadística: Madrid.

Jones C (2017) Spatial economy and the geography of functional economic areas. *Environment and Planning B: Urban Analytics and City Science*, *44*(3), 486-503.

Jones C, Coombes M and Wong C (2012) A system of national tiered housing-market areas and spatial

planning. *Environment and Planning B: Planning and Design*, *39*(3), 518-532.

Kim H, Chun Y and Kim K (2015) Delimitation of Functional Regions Using a p-Regions Problem Approach. *International Regional Science Review*, *38*(3), 235-263.

Lambiotte R (2010) Multi-scale modularity in complex networks. In *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks* (pp. 546-553). IEEE.

Lancichinetti A and Fortunato S (2011) Limits of modularity maximization in community detection. *Physical review E*, *84*(6), 066122.

Masser I and Scheurwater J (1980) Functional regionalisation of spatial interaction data: an evaluation of some suggested strategies. *Environment and Planning A: Economy and Space*, 12, 1357-1382.

Martínez-Bernabéu L and Casado-Díaz JM (2016) Delineating zones to increase geographical detail in individual response data files: An application to the Spanish 2011 Census of population. *Moravian Geographical Reports*, *24*(2), 26-36.

Martínez-Bernabéu L, Coombes M and Casado-Díaz JM (2020) Functional Regions for Policy: a Statistical 'Toolbox' Providing Evidence for Decisions between Alternative Geographies. *Applied Spatial Analysis and Policy*, *13*(3), 739-758.

Martínez-Bernabéu L, Flórez-Revuelta F and Casado-Díaz JM (2012) Grouping genetic operators for the delineation of functional areas based on spatial interaction. *Expert Systems with Applications*, *39*(8), 6754-6766.

Masser I and Brown PJ (1975) Hierarchical aggregation procedures for interaction data. *Environment and Planning A: Economy and Space*, *7*(5), 509-523.

Mu X and Yeh AGO (2020) Regional delineation of China based on commuting flows. *Environment and Planning A: Economy and Space*, *52*(3), 478-482.

OECD (2002) *Redefining Territories. The Functional Regions*. Paris. OECD Publications.

OECD (2012) *Redefining Urban: a new way to measure metropolitan areas*. Paris. OECD Publications.

R Core Team (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ (accessed 1 Jun 2020).

Shen Y and Batty M (2019) Delineating the perceived functional regions of London from commuting flows. *Environment and Planning A: Economy and Space*, *51*(3), 547-550.

United Nations Statistics Division (2019) *2020 World Population and Housing Census Programme*. Available at: https://unstats.un.org/unsd/demographic-social/census/ (accessed 1 Jun 2020).

Wicht A, Kropp P and Schwengler B (2020) Are functional regions more homogeneous than administrative regions? A test using hierarchical linear models. *Papers in Regional Science*, *99*(1), 135-164.