Istat implementation of the algorithm to develop Labour Market Areas

Technical Report

Eurostat Grant on “EU-TTWA method: improvements, documentation and sharing knowledge activities”

Luisa Franconi[ǂ], Michele D’alò[ʄ] and Daniela Ichim[ǂ]

13 June 2016, v.1.0

Table of content

[ǂ] Istat, Italian National Statistical Institute, Directorate for territorial and environmental statistics, franconi@istat.it

[ʄ] Istat, Italian National Statistical Institute, Directorate for methodology and statistical process design, dalo@istat.it

[ǂ] Istat, Italian National Statistical Institute, Department for statistical production, ichim@istat.it

1. Introduction

Labour market areas (LMAs) are sub-regional geographical areas where the bulk of the labour force lives and works, and where establishments can find the main part of the labour force necessary to occupy the offered jobs.

LMAs represent a functional geography i.e. geography based on relationships between elementary territorial units: these can be municipalities, census output areas, provinces, etc. In what follows we will call such elementary unit a community. A LMA is an aggregation of communities; such aggregation is made in order to maximize the internal relationships amongst communities and should satisfy criteria defined by the user. The relationships are represented by commuting flows.

This report intends to provide a detailed description of the algorithm implemented in R (R Development Core Team, 2011) named LabourMarketAreas that has been initially presented in Coombes and Bond (2008) and briefly sketched in the Final report of the Eurostat Task Force on Harmonised Labour Market Areas (Eurostat, 2015).

This method represents an evolution of the classical methodology of the "*Travel-To-Work-Areas*", defined in Coombes *et al*. (1986), and adopted, with variations in several different countries (see Casado Díaz and Coombes, 2011 for comparisons amongst them and Coombes et al., 2012 for further analysis on possible harmonization at European level). The initial TTWA algorithm has been used also in Italy (Istat, 1997 and Istat and IRPET, 1989) to create the LMAs for the years 1981, 1991 and 2001. The 2011 version of the Italian LMAs has been developed using the algorithm described in this report (Istat, 2014). This description is based on a more extended version in Italian (Istat, 2015 Chapter 1, section 2).

In Section 2 the definition of LMAs is provided together with the notation used. Section 3 provides the description of the algorithm and Section 4 presents its flow chart.

2. Labour Market Areas

LMAs are clusters comprising two or more communities. Such clusters to be called Labour Market Areas should satisfy a set of principles already outlined in Eurostat (1992) and recognized in the literature as of primary importance comprising. These principles comprise objectives of LMA (to be statistically defined areas each representing a labour market); constraints (being a partition of the whole country, each part formed by contiguous communities); criteria (autonomy, homogeneity, coherence and conformity) and finally the need to be flexible in order to cope with elementary units that could be of very different sizes.

The most distinctive characteristic of a labour market area is the ability to maximise the relationships inside its border and minimise them across borders. The way to quantify this quality is by means of the concept of self-containment of incoming and outgoing flows.

Let f_{hk} be the flow between community (or group of communities) h and community (or group of communities) k, i.e. the number of commuters living in h and working in k. Then

$R_i = \sum_k f_{ik}$ is the number of employees[1] living in area i, $W_i = \sum_h f_{hi}$ is the number of employees working in area i and $RW_i = f_{ii}$ is the number of employees living and working in area i. There are two types of self-containment: the supply side self-containment defined as $SS\_SC = RW_i / R_i$ and the demand side self-containment defined as $DS\_SC = RW_i / W_i$. These two quantities measure the level of internal cohesion or integration of the areas with respect to the commuting flows.

3. The implementation: a technical description

3.1 An overview

The aim of the algorithm is to determine a partition of a set of communities based on the horizontal relationships between them. Such relationships take the form of commuting flows between communities usually recorded by the population census or taken from administrative registers.

The algorithm is an iterative agglomerative algorithm that depends on a set of parameters. These parameters set the level of desired self-containment and size of the LMAs. It starts by considering each community as a cluster that is checked against a set of conditions to see whether it can be considered an LMA. At each iteration clusters that are not fit for the purpose are disaggregated and a single community inside the cluster is chosen to be attached to a new cluster that improve the set of given conditions. The final solution is obtained when the whole set of clusters satisfies the given conditions. What is needed then is: a set of parameters, a function to decide when a cluster is “fit for the purpose”, a measure to choose to which cluster assign the selected community and the steps of the iterative procedure. These elements are described in the next section.

3.2 The components of the algorithm

The algorithm is based on the following components:

1. A set of parameters, chosen by the user, that identifies thresholds on the size of the LMA, in terms of number of residents, and on the level of self-containment required in order for a cluster to be considered an LMA;
2. A condition of validity that, based on the values of the parameters, establishes the criteria that should be met by a cluster in order to be considered an LMA and quantifies whether the identified cluster is a valid LMA;
3. A measure of cohesion between a community and all the clusters with whom such community has relationships; such measure identifies the cluster where the community will to be assigned: the one where the maximum is attained. Currently if two or more clusters share this value, the first one is taken;

[1] In this report we use the term employees to indicate the working people who are defined eligible as commuters by the national census or by the actual source of the data. Different countries adopt slightly different definitions.

4. A reserve list (Coombes, 2014) comprising of communities which cannot be clearly assigned during the iterations of the algorithm;
5. An iterative procedure that selects a community at a time and aggregates it to a different cluster and defines the operations to be implemented.

2.2.1 The parameters of the algorithm

Usually in TTWA methods the parameters are two: one related to the minimum size of a cluster to consider it an LMA and the other related to the self-containment (either one of the two self-containment defined in Section 2.1, usually the demand side self-containment or the minimum between the two) necessary to determine the minimum level of internal relationship.

Coombes and Bond (2008) allows for a degree of flexibility by defining four parameters in order to introduce a trade-off between the number of occupied people residing and an area (its size) and the level of integration needed to consider this area a LMA (its level of self-containment). In particular smaller areas will need a higher level of self-containment than larger areas in order to be considered LMAs. The parameters are summarized in Table 1.

Table 1: The algorithm parameters and their meaning.

Parameter	Meaning
minSZ	minimum number of employees for a cluster to be considered an LMA
tarSZ	target value for the size of the cluster i.e. the value for which we can accept a lower level of self-containment for an LMA
minSC	level of self-containment that is acceptable for cluster of large sizes
tarSC	the minimum level acceptable for the minimum self-containment SC, $SC = \min(SS\_SC, DS\_SC)$ in order for a small cluster of communities to be considered an LMA

The parameter tarSC is always greater than minSC. Common values for tarSC are between 0.75 and 0.8 but in specific situation also higher values can be selected. For the parameter minSC possible values are usually between 0.6 and 0.7.

The reasoning behind these parameters related to self-containment is linked to the idea of the "number of employees staying in the LMAs": 0.75 means in general terms that three out of four employees stay inside the area, 0.6667 corresponds to four out of six, 0.6 three out of five and so on.

As far as the parameters of the size are concerned there are different aspects to be considered: the minimum size answers to the question "what is the size of the smallest labour market that I can consider?" whereas the target size rally depends on the data. Usually it takes value from 10 thousands onwards.

2.2.2 The validity condition

The validity condition states whether a cluster is a proper LMA or not. The condition is operatively defined through a function that expresses the trade-off between dimension (in terms of occupied persons), SZ, and the self-containment, SC, of the cluster. This validity function, f_v, depends also on the selected parameters (described in section 2.2.1) and takes the following form:

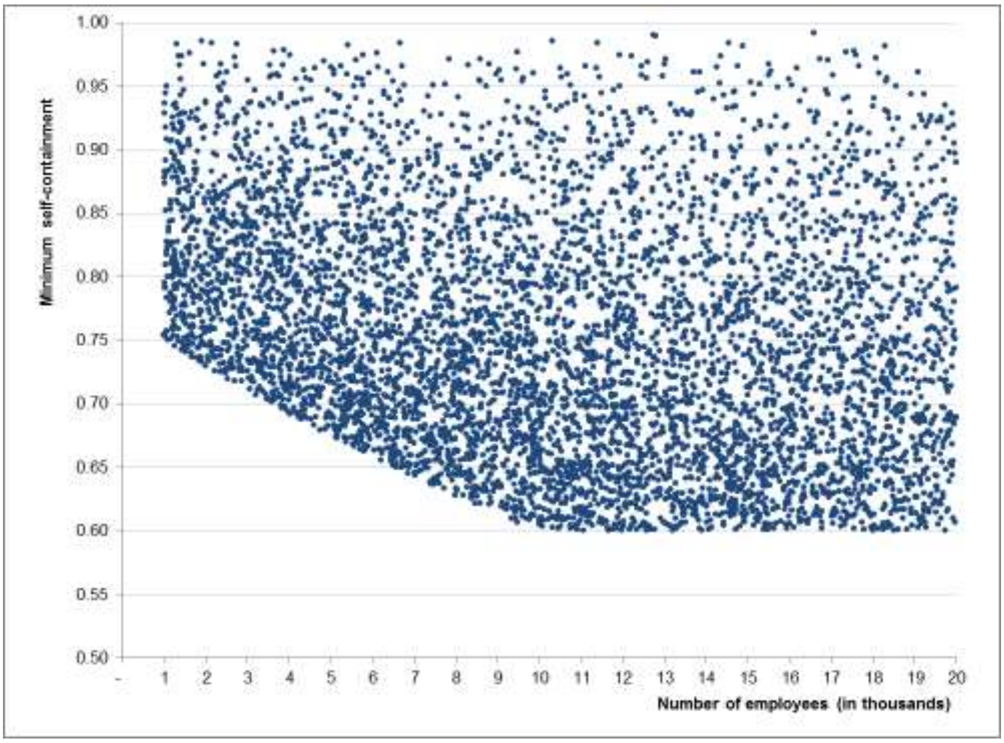
$$f_v(SZ, SC) = \left[1 - \left(1 - \frac{minSC}{tarSC}\right) \cdot \max\left(\frac{tarSZ - SZ}{tarSZ - minSZ}, 0\right)\right]\left[\frac{\min(SC, tarSC)}{tarSC}\right] \qquad (1)$$

The validity condition states that a cluster with size SZ_c and self-containment SC_c (minimum between SS_SC_c andDS_SC_c) is a proper LMA if:

$$f_v(SZ_c, SC_c) \geq \frac{minSC}{tarSC}. \qquad (2)$$

This condition is therefore evaluated at each iteration to check whether all the clusters are indeed proper LMAs.

Figure 1: Graphical representation of the validity function: each point in the graph is a combination of size and self-containment which corresponds to an acceptable LMA.



2.2.3 The cohesion measure

The algorithm is an aggregative one therefore it is necessary to define a measure that establishes to which cluster the selected community ought to be aggregated in order to maximize the interaction of the communities inside the cluster.

The measure of interaction between community *h* and cluster *k*, L_{hk} , (Smart, 1974) is called cohesion measure. It is defined by means of the sums of inflow and outflow between the involved entities standardized with respect to the relative origin and destination and takes the form:

$$L_{hk} = \left[\frac{(f_{hk})^2}{(R_h W_k)}\right] + \left[\frac{(f_{kh})^2}{(R_k W_h)}\right] . \tag{3}$$

This measure implements the concept of reciprocal importance, i.e. it divides the attraction capacity of the entity by the total incoming flows.

The cluster that maximizes the cohesion for the selected community *h* is called the dominant cluster for *h* and *h* is assigned to it.

2.2.4 The reserve list

The reserve list comprises all communities that during the iterative process of allocation did not bring to an improvement. In particular, either the assignment of these communities to a cluster does not improve its validity or a dominant cluster for them does not exist.

The communities in such list are taken away from the analysis and assigned to a cluster only at the end of the process when all the clusters satisfy the validity condition. Such list is intended as an extra flexibility of the algorithm.

To allow for further investigation of the characteristics of the communities belonging to the reserve list some information are recoded by the R package LabourMarketAreas. The iteration in which the assignment to the reserve list has taken place, the value of the validity, the reason for the assignment: lack of dominant cluster or failure in improving the validity of the dominant cluster. It is also recorded whether the cluster to which the community belongs is formed just by this community (single cluster) or if it comprises more communities. The various combinations of these two factors give rise to the classification reported in Table 2.

Table 2: Classification of assignment to the reserve list for a community by reason and number of communities of the initial cluster.

Type	Reason	# Communities
A	Failure in improving the validity of dominant cluster	1
B	Dominant cluster does not exists	1
C	Failure in improving the validity of dominant cluster	More than one
D	Dominant cluster does not exists	More than one
E	If the originally selected community has no dominant cluster, then all the communities in the cluster are assigned sequentially. This type of assignment to the reserve list occurs when one of the other communities in such cluster does not increase the value of the validity	More than one
F	If the originally selected community has no dominant cluster, then all the communities in the cluster are assigned sequentially. This type of assignment to the reserve list occurs when one of the other communities in such cluster does not have a dominant cluster.	More than one

2.2.5 The iterative procedure

The iterative procedure is composed by the following steps:

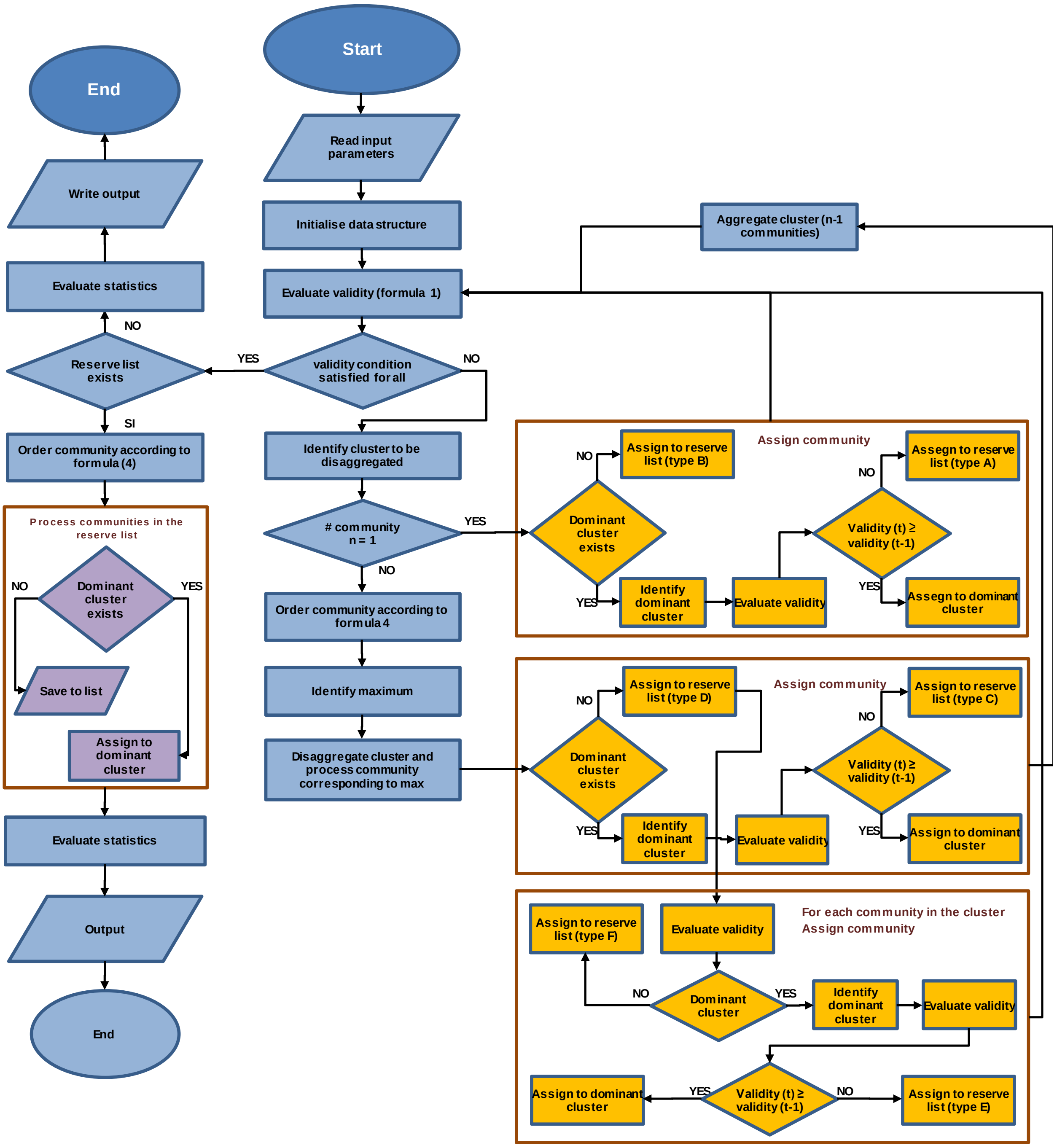
1. Each initial community is considered a cluster. For each of them the validity function (1) is evaluated;
2. Until all clusters do not satisfy the condition (2) repeat:
    a. Identify the cluster *S* with minimum validity and disaggregate it into constituent communities;
    b. Among the communities of cluster S identify the community *h* that maximises external relationships i.e. maximises the flows, $ord(h)$, with other clusters:
    $ord(h) = \sum_{i \notin S, h \in S} f_{ih} + \sum_{h \in S, j \notin S} f_{hj}$ (4)
    This community is selected to be aggregated to a new cluster.
    c. Identify the dominant cluster D for community *h* by means of the cohesion measure L_{hk};
    d. If the dominant cluster D exists, a check is needed on the value of its validity. If the addition of community *h* to cluster D increases its validity with respect to the initial composition then the community *h* is assigned to cluster D. If, on the contrary, the new cluster validity is not greater than the original value, the community *h* is taken away from the list of communities to be analysed from the algorithm and it is assigned to the reserve list. This assignment to the reserve list is called of type A or type C depending on the number of communities in the cluster: respectively one or more than one;
    e. If the dominant cluster does not exist, the community h is assigned directly to the reserve list (type B or type D depending on the number of communities in the cluster: respectively one or more than one). In case the cluster S to which the community h belongs is formed by more than one municipality, all such communities are assigned one after the other, according to the order established by ord to the corresponding dominant clusters if they exist or, alternatively, to the reserve list either if the dominant cluster does not exists (type F), or if the value of the validity does not increase (type E);
3. Evaluate the validity of the new set of clusters and return to step 2. above.

The algorithm repeats the same combination of disaggregation and aggregation till convergence to the final solution where all the clusters satisfy the validity constraint.

Once the solution is determined all the communities in the reserve list are assigned to the dominant cluster or, if it does not exist, they are reported in a special list.

4. Flow chart of the algorithm

Figure 2: Flow chart of the algorithm.


Start
Read input parameters
Initialise data structure
Evaluate validity (formula 1)
validity condition satisfied for all
YES
NO
Identify cluster to be disaggregated
# community n = 1
YES
NO
Order community according to formula 4
Identify maximum
Disaggregate cluster and process community corresponding to max
End
Write output
Evaluate statistics
NO
Reserve list exists
SI
Order community according to formula (4)
Process communities in the reserve list
NO
Dominant cluster exists
YES
Save to list
Assign to dominant cluster
Evaluate statistics
Output
End
Aggregate cluster (n-1 communities)
Assign community
NO
Assign to reserve list (type B)
Dominant cluster exists
YES
Identify dominant cluster
Evaluate validity
Validity (t) ≥ validity (t-1)
NO
Assegn to reserve list (type A)
YES
Assegn to dominant cluster
Assign community
NO
Assign to reserve list (type D)
Dominant cluster exists
YES
Identify dominant cluster
Evaluate validity
Validity (t) ≥ validity (t-1)
NO
Assign to reserve list (type C)
YES
Assign to dominant cluster
For each community in the cluster Assign community
Assign to reserve list (type F)
Evaluate validity
NO
Dominant cluster
YES
Identify dominant cluster
Evaluate validity
Assign to dominant cluster
YES
Validity (t) ≥ validity (t-1)
NO
Assign to reserve list (type E)

5 References

Casado Díaz, J., 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, pp. 7–32.

Coombes, M. (2014). Personal communication to Task Force members.

Coombes, M., Bond, S. (2008). Travel-to-Work Areas: the 2007 review. London: Office for National Statistics, 2008. (http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/other/travel-to-work-areas/index.htm)

Coombes, M., Casado-Diaz, J.M., Martìnez-Bernabeu, L., Carausu, F. (2012). Study on comparable labour market areas: final research report. 17 October 2012. Eurostat-Framework contract nº :6001. 2008.001 - 2009.065, Specific contract nº: 50405.2010.004 – 2011.325.

Coombes M.G., Green A.E., Openshow S., (1986). An efficient algorithm to generate official statistics report areas: the case of the 1984 Travel-to-Work Areas in Britain. *The Journal of Operational Research Society.* 37(10): 943–953.

Eurostat (2015). Final report of Eurostat Task Force on “Harmonised Labour Market Areas”. Available at: http://ec.europa.eu/eurostat/cros/system/files/Task%20Force%20on%20LMA%20Final%20Report.pdf_en

EUROSTAT (1992) Study on employment zones (E/LOC/20) Office for Official Publications of the European Communities, Luxembourg.

Istat (1997). *I Sistemi Locali del Lavoro 1991*. Pagg. 235–242. Istituto Poligrafico e Zecca dello Stato. Roma.

Istat and IRPET (1989). *I Mercati Locali del Lavoro*. Franco Angeli. Milano.

Istat (2014). Labour Market Areas Year 2011. Roma. 17 dicembre 2014. Available at: http://www.istat.it/en/files/2014/12/EN_Labour-market-areas_2011.pdf?title=Labour+Market+Areas+-+17+Dec+2014+-+Full+text.pdf

Istat (2015). *La nuova geografia dei sistemi locali*. Letture statistiche - Territorio. E-book: http://www.istat.it/it/archivio/172444.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Smart**,** M. W. (1974). Labour market areas: uses and definition. *Progress in Planning* 2 (4), 239–353.