**Building the territorial statistical register: quality control on geocoded administrative data**

Daniel Ibáñez Vidal, Institut d'Estadística de Catalunya, dibanez@idescat.cat

Mònica Gasulla Ramon, Institut d'Estadística de Catalunya, mgasulla@idescat.cat

Eduard Suñé Luís, Institut d'Estadística de Catalunya, esl@idescat.cat

Mireia Farré Mallofré, Institut d'Estadística de Catalunya, mfarre@idescat.cat

Cristina Rovira Trepat, Institut d'Estadística de Catalunya, crovira@idescat.cat

Frederic Udina i Abelló, Institut d'Estadística de Catalunya, fudina@idescat.cat

**Abstract**

*The Statistical Institute of Catalonia (Idescat) is developing a statistical information system based on administrative registers. All these registers usually contain a set of variables defining the spatial location of the microdata through a postal address. One of the subsystems used is the information about real estate cadastre, which contains around 10 million georeferenced estates. Our aim is to analyse its statistical quality.*

*A set of polygons was created, associated with the concave or convex hull from the centroids of the estates under the same road code, as a first step in establishing the quality of the positions. One of the tasks carried out is that of the detection of outliers in these polygons. We focus on analysing the distribution of the distances between the points of each polygon.*

*First, we create an indicator to rank the polygons according to their level of reliability, indicating the possibility of there being an outlier in the polygon. Our first concern is to identify the polygons with the most extreme values on the indicator, both those which could contain suspicious points (around 3%) and those do not contain any (approximately 90%).*

*Secondly, our concern is to study the cases found in the intermediate zone, and in particular to check in which threshold we will obtain the optimal relation between the well-classified and the badly-classified ones. In order to evaluate our indicator, we must select a sample of cases and study them manually to ascertain whether they are correct or not.*

**Keywords:** Statistical territory register, Real estate register, georeferencing, statistical quality, outliers

# Purpose

The purpose of the study is to define a method to control the quality of the georeferenced information within the statistical production process, in accordance with the Generic Statistics Business Process Model (GSBPM).

Within this context, in order to analyse the quality of the geolocation of property (buildings and dwellings) in Catalonia, indicators have been created which can be used to detect the presence of anomalous points in a set of polygons representing each road in the territory. These polygons have been defined upon the basis of the centroids of the plots of land where the property is located. A point is regarded as anomalous when its distance with respect to the other points on the same road is suspiciously high, generally due to an error in the assignment of the road. We will regard a polygon as erroneous when it contains at least one anomalous point.

# Context

## Statistical registration of territory

Idescat, the Statistical Institute of Catalonia, is constructing a production system, chiefly based on administrative records made up of three different subsystems: the Statistical Population Register (REP), the Statistical Territory Register (RET) and the Statistical Entity Register (REE).

The **Statistical Territory Register** (RET) is a spatial information subsystem whose main purpose is the geolocation and validation of the postal addresses associated with the different statistical units which appear in the other subsystems mentioned above (population and entities).

The postal addresses contained in the RET, provided by the National Statistics Institute (INE), are geocoded by means of a web service belonging to the Cartographic and Geological Institute of Catalonia (ICGC). In addition, the complete postal addresses must be validated to check that they match the information on property provided by the General Directorate of Cadastre (DGC).

The comparison between the postal addresses of the INE and the DGC entails difficulties, due to the lack of consistency in the types, names and codes of the roads. For this reason it has been necessary to create a correlation table between the two sources. To compile the above table, criteria of similarity between literals and the relationships between the geometrical representations of the roads have been used.

## Geometrical representation of the roads: construction of polygons

The General Directorate of Cadastre (DGC) has administrative records on all the property in the territory. They specify the postal address of each property and its link to the plot of land to which it corresponds, by means of which the position of the centroid is identified. By grouping together all the centroids of the plots of land for the same road, a polygon can be constructed to represent the geometry of the road. These polygons are constructed by means of the operation known as “concave hull” and, when this does not provide a solution, the “convex hull” (Asaeedi, Didehvar and Mohades, 2017).

Similarly, another set of polygons is obtained by using the geocoded INE addresses.

During the process to compare the two sets it was observed that some DGC polygons contained anomalous points, as a result of which the need to establish information quality indicators was detected.

## Database creation process

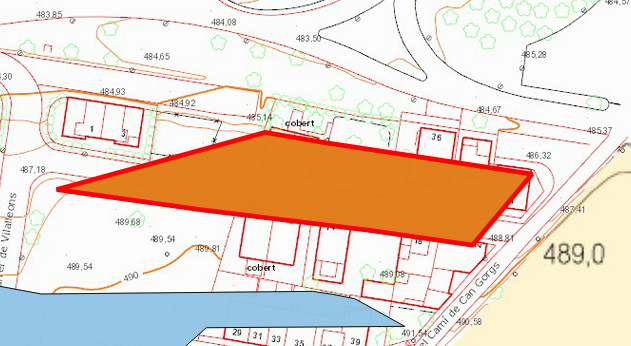
With the DGC’s information on each of the roads (identifier, type of road, name and area), an initial database with 81,240 polygons was created, representing them graphically and biunivocally.

These polygons are made up of a set of **vertices** (points) which, in turn, determine each of their **sides**. With the lengths of the 882,000 sides associated with each polygon, a second database was constructed to generate a battery of variables for the polygons, such as the sum total of the lengths of the sides (perimeter) and the average length of the sides.

**Side with the second greatest length**

Figure . Example of a polygon with four vertices.

**Vertex**



**Side with the shortest length**

**Side with the greatest length**

Source: Compiled by the authors with information from the General Directorate of Cadastre

In addition, variables were constructed upon the basis of the differences between the lengths of the sides. To do so, the sides were sorted in accordance with their decreasing length. The difference between the lengths was then calculated in accordance with the previous sorting criterion. More variables for the polygons were created with these differences, such as the sum total of these differences, their average and the maximum. All the calculated variables are associated with their corresponding polygon in the polygon database to be covered by the study.

The **type of road** has an effect which may condition the analysis, as each type of road is associated with a very different polygon standard: a polygon which represents a depopulated area is not the same as that of a lane or a path. To prevent this potential effect, it was decided to exclude the polygons representing certain types of particularly singular roads from the study, such as outskirts or depopulated areas. As a result, **the work database contains 79,344 polygons**.

# Analysis process

Determining whether there is a side oddly superior to the others in each polygon in order to classify it as erroneous is a problem which prevents us from establishing a determinist criterion for its solution. Supervised methods have been chosen from among the statistical classification methods, due to the relative ease with which erroneous polygons can be visually identified.

These methods require having at least two correctly labelled data sets containing a high number of erroneous and valid cases, one for the training or calculation phase and the other for testing and validating the effectiveness of the algorithm or classifier resulting from the initial phase.

1,150 polygons were analysed, guaranteeing a sufficient presence of anomalous polygons which could be labelled as correct or erroneous. Of the above, 900 were randomly assigned to the training set and 250 to the test set. 30.3% of each of these sets are anomalous polygons.

For the univariate methods, each classifier is obtained by establishing a threshold of the variable above which a polygon will be labelled as erroneous. In this way, a variable or indicator is interpreted as a function of the likelihood of a polygon containing an anomalous point.

## Step 1: univariate study

The discriminant capacity of each of the work variables was evaluated by comparing the Area Under the Receiver Operating Characteristic curve (AUROC)[[1]](#footnote-2).

This first study warned of the behaviour of one of the variables: **the length of the longest side.** The use of this variable was sometimes unnecessary and even counterproductive. In most of the cases of erroneous polygons there are a minimum of two sides with lengths longer than the others (see figure 2), as a result of which, by excluding the side with the maximum length, the indicator continues to have a similar value.

Figure . Example in which an erroneous point (A) entails having two sides with lengths longer than the others.

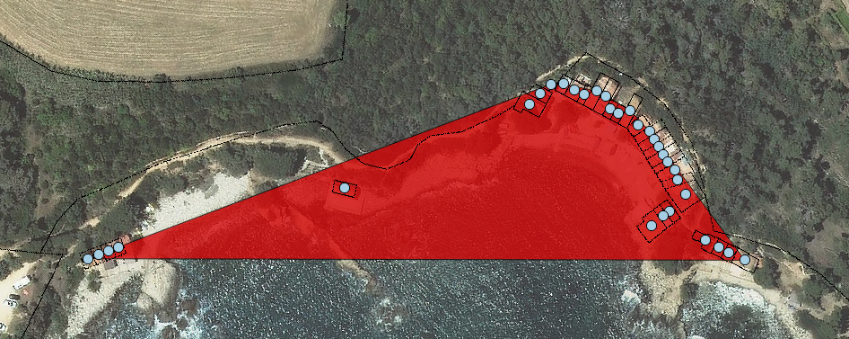


**Point A**

Source: Compiled by the authors with information from the General Directorate of Cadastre

Conversely, in some types of roads, such as roads with a single pavement, the furthest vertex has only one associated side with a length greater than the others (see figure 3), as a result of which, by excluding the side with the greatest length, this case will no longer count as a potential error (and will therefore be a correct prediction).

Figure . Example in which the side with the greatest length is not a symptom of an erroneous polygon.



Source: Compiled by the authors with information from the General Directorate of Cadastre

As a result, **an additional database** was built, **consisting of the lengths of each side of a polygon, except for the maximum length.** All the variables for the polygons were recalculated and added to the work database.

Taking into account all the variables defined (out of a total of 20), it was observed that many of them proved to be good classifiers. These variables include the maximum difference between the lengths (of the database of polygons without the longest side), their sum total, the second maximum length and the standard deviation of the lengths. Conversely, the number of vertices of the polygon and the length of the shortest side did not predict the response very well on an individual basis.

## Step 2: use of standard multivariate models

In order to obtain higher sensitivities, the following multivariate algorithms were executed, frequently used in the field of supervised classification: logistic regression, conditional inference trees and discriminant analysis[[2]](#footnote-3).

These methods did not provide the expected optimal results, as outlined in the results section (see Table 1), although they did help to identify which set of variables displayed the most discriminatory power, enabling us to modify and optimize some of the subsequently designed indicators.

## Step 3: proposal of indicators

It was decided to create univariate indicators by means of the combination of database variables, due to the poor results of the previously tested models. Most of these indicators were constructed by using non-linear combinations of the variables which displayed the best results in the previous analyses.

A threshold providing 90% sensitivity in the training set was calculated for each of these indicators. When this threshold was transferred to the test set, only the indicators which displayed a sensitivity similar to that established for the training set were maintained.

At this point the main purpose of the analysis had been achieved, namely to obtain a satisfactory instrument for the detection of anomalous polygons. However, it was decided to establish an additional objective: to distinguish the subset of polygons with the greatest likelihood of containing errors among those classified as anomalous. This would permit the setting of priorities in the clean-up exercise.

The following conditions were established to select an indicator in keeping with this objective:

1. A threshold guaranteeing 95% specificity was chosen for the group of anomalous polygons.
2. To guarantee the consistency of the prevalence in the training and test sets.

The indicator thus selected, **the polygon quality indicator (PQI),** combines the **maximum difference** and the **variation coefficient** of the lengths of its sides. Both variables refer to the **polygon without taking into account the longest side** (see 2.3). This indicator sorts the polygons according to their likelihood of containing anomalous points.

It is expressed for each polygon as follows:

(1)

Where “” iterates along the sides of the polygon according to its decreasing length, “” is its number of vertices and is the variation coefficient of the lengths of the sides:

(2)

This indicator, in addition to classifying and sorting the polygons in accordance with their estimated quality, allows the calculation of the percentage of polygons liable to be erroneous (GQ1) and the percentage of polygons requiring priority revision (GQ2).

# Results

The first of the result tables compares the sensitivities of the models tested with the polygon quality indicator (PQI).

**Table 1. Comparison of the results of the different methods in the test set (250 cases)**

|  |  |  |
| --- | --- | --- |
|  | **Polygons calculated to be erroneous** | **Sensitivity** |
| PQI | 141 (56.4%) | 89.0% |
| Conditional inference tree | 66 (26.4%) | 57.3% |
| Logistic regression | 53 (21.2%) | 52.8% |
| Quadratic discriminant analysis | 48 (19.2%) | 46.1% |
| Linear discriminant analysis | 21 (8.4%) | 22.5% |

Source: Compiled by the authors

The PQI summary for the training set, the test set and the population base is shown in table 2. It should be borne in mind that the sets were selected to guarantee a broad presence of anomalous polygons.

**Table 2. Statistical summary of the PQI**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Minimum** | **Q1** | **Average** | **Median** | **Q3** | **Maximum** | **Variance** |
| Training set (n=900) | 0.0 | 12.5 | 80.9 | 1,316.0 | 283.3 | 60,560.0 | 5,111.5 |
| Test set (n=250) | 0.0 | 16.7 | 81.6 | 1,226.0 | 297.4 | 38,720.0 | 4,495.2 |
| Population base (n=79,344) | 0.0 | 1.1 | 3.6 | 189.9 | 12.8 | 132,600.0 | 2,186.3 |

Source: Compiled by the authors

The prevalence of erroneous polygons of the PQI for the groups of polygons determined by two different thresholds can be observed below (table 3):

* QG1 - Erroneous polygons: 90% sensitivity
* QG2 - Polygons with a very high likelihood of being erroneous: 95% specificity, QG2 ⊆ QG1.

The distribution of the polygons according to these groups in the population base is also shown. GQ2 (which includes the polygons with the greatest likelihood of being erroneous) only contains 2.9% of the total number of polygons.

**Table 3. Prevalence of erroneous polygons of each group established by the PQI and its distribution in the population base**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Prevalence of erroneous polygons** | |  |  | **% polygons in the database** |
| **Training set** | **Test set** |
| GQ1 | 48,2 | 45,1 | GQ1 | 8,9 |
| GQ2 | 81,0 | 90,9 | GQ2 | 2,9 |

Source: Compiled by the authors

To end, by way of example, to illustrate the usefulness of the PQI indicator, in cartogram 1 we can visualize the results obtained by means of the scores of the polygon quality indicator in this classification. The polygons of this cartogram are represented by different tonalities, depending on the value of the indicator: darker means a greater likelihood of an error and vice versa.

Cartogram 1. PQI of a Catalan municipality (Vilanova i la Geltrú).



Source: Compiled by the authors with information from the General Directorate of Cadastre

# Conclusions

In order to confirm the quality of the spatial information provided by the General Directorate of Cadastre (DGC) by means of an explanatory analysis, the construction of a quality indicator was proposed. This facilitates the detection of inconsistencies between the road identifiers and the centroids of the associated properties. To do so, a set of polygons was constructed to geometrically represent all the roads, revealing the existence of potential erroneous polygons.

A set of variables related to the lengths of the sides was defined in order to estimate the likelihood of a polygon being erroneous. To this effect we used multivariate supervised classification methods, which did not provide us with satisfactory results.

We therefore opted for the creation of a set of indicators, most of them non-linear, based on the calculated variables. Firstly, the indicators producing the best classifiers were selected by means of AUROC methods.

In addition, it was decided to establish a further objective: to distinguish the subset with the greatest likelihood of containing errors among those classified as erroneous.

The indicator selected, **the polygon quality indicator (PQI)**, combines the **maximum difference** and the **variation coefficient** of the lengths of its sides. Both variables refer to the **polygon without taking into account the longest side**.

This indicator not only classifies and identifies the polygons in accordance with their estimated quality, it also enables us to calculate the percentage of polygons liable to be erroneous (GQ1) and the percentage of polygons requiring priority revision (GQ2).

# References

Asaeedi, S., Didehvar, F. and Mohades, A. (2017). α-Concave hull, a generalization of convex hull. Theoretical Computer Science, 702, pp.48-59.

Chen, D., Lu, C., Kou, Y. and Chen, F. (2007). On Detecting Spatial Outliers. GeoInformatica, 12(4), pp.455-475.

Huynh H.T., Hoang M.T., Vo N.H., and Won Y. (2006), Outlier Detection with Two-Stage Area-Descent Method for Linear Regression, Proceedings of the 6th WSEAS International Conference on Applied Computer Science, Tenerife, Canary Islands, Spain, December 16-18, 2006

Lalkhen, A. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain, 8(6), pp.221-223. (https://doi.org/10.1093/bjaceaccp/mkn041)

Molnar, C. (2013). Recursive Partitioning by Conditional Inference. Available online at: <https://github.com/christophM/overview-ctrees> (Accessed: 22 February 2018).

1. The ROC curve is a chart which compares sensitivity (true positives divided between true positives and false negatives) and specificity (true negatives divided between true negatives and false positives) (Lalkhen and McCluskey, 2008). In this study a polygon labelled as erroneous and which actually is erroneous is regarded as a true positive. [↑](#footnote-ref-2)
2. It is not possible to provide a more detailed description of these methods due to space restrictions in the format of this communication [↑](#footnote-ref-3)