# Automatic Delineation of Polling Station Areas using Legal Texts

Didier Josselin
UMR ESPACE 7300, CNRS, LIA
Avignon, France
didier.josselin@univ-avignon.fr

Soumaya Yahiaoui
Université d'Avignon, LIA
Avignon, France
yahiaoui.soumaya@gmail.com

Eric San Juan
Université d'Avignon, LIA
Avignon, France
eric.sanjuan@univ-avignon.fr

Christèle Marchand-Lagier
Université d'Avignon, LBNC
Avignon, France
christele.marchand@univ-avignon.fr

**Abstract**

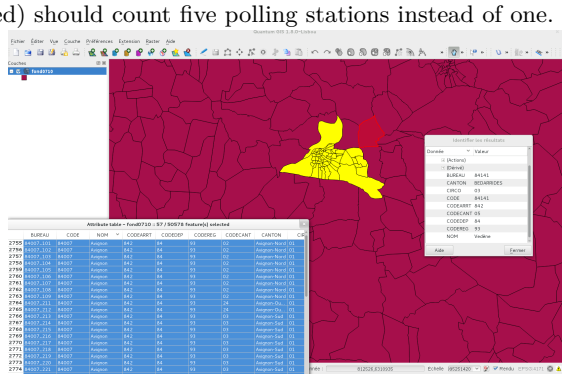
Using free GIS (QGIS/POSTGIS), regular expression processing and geographical data for networks and addresses, we propose an automatic computation to verify and to (re)draw polling area boundaries from official legal texts. Indeed, middle size French towns miss this important geographical information. Contributions and limitations of the methods used and the results obtained in this work are presented and discussed. Perspectives are also drawn.
*Keywords*: geographical information retrieval, regular expressions, polling station boundary, geometry, topology

## 1 Electoral partitioning context

Statistical information on territories does not always perfectly match the corresponding spatial entities due to possible erroneous geocoding. Since direct universal suffering in France in 1848, territorial partitioning provides an interesting information about the relation between population and policy [9, 7]. It is indeed easy to find data at different administrative scales. Those are sometimes very costly to generate and can suffer from a lack of quality (accuracy, completeness, missing metadata, semantic ambiguity). Especially, a part of this information is sometimes written in texts and there does not exist any complete and accurate map to ensure the data location reliability. It is indeed the case for electoral roll information. A voter is identified according to his/her address and assigned to a unique polling station. However, at least in France, there is no updated spatial complete partition of the polling stations, because is it expensive to build and it may change from an election to another. For some municipalities, the only information we can handle is a series of legal texts, which somehow «tell» us the polling station boundary structure. Combining numerical geographical information and these reference texts is the problem tackled in this article. The final aim is to draw reliable, accurate, complete and topologically correct boundaries of the polling station polygons.

In France, about 30000 municipalities are so little that they only have a single polling station. For those and for the largest towns and cities (Paris, Lyon, Nantes...) where urban administration offices have digitized the polling station partition, there is no major problem, except that boundaries are sometimes uncertain and marred by errors. Those are all listed in the Cartelec maps [10] (cf. Figure 1). This database was populated using a semi-automated process that is difficult to maintain on the long term, because it is based on paper map digitizing. Moreover, there are still about 5000 municipalities composed of several stations, which cannot afford spatial data digitization. Whatever the type and the size of municipalities, they also may modify their electoral partition by merging, splitting or changing the boundaries, which would lead to an obsolete cartography.

Figure 1: The French national CARTELEC cartography of polling stations: a few are well delineated in Avignon (yellow), but the municipality of Vedène (North-East, in red) should count five polling stations instead of one.



On the other hand, legal texts are supposed to be reliable and effective, as the reference documents. A sensitive constraint to respect is to ensure a complete spatial partition in such a way that any person can vote in only one station. The French law imposes each station to have a «geographical perimeter». Due to a possible risk of election invalidation, it is recommended the electoral partition to have no error [1]. The associated reference legal texts are public but sometimes difficult to find. They are

most of the time made of regular expressions, according to a specific legal thesaurus. Thence, we can rely on these text to build geometries because they include spatial information [20, 19].

Therefore, an automatic and generic method should be developed to delineate polling stations areas from reference legal texts for French middle size towns.

## 2 Aims of the research

To build a reliable and relevant partition of polling station areas, we first propose to test the process on simple and well designed legal texts, which as much as possible respect the associated polygonal geometries. The partition should be accurate and should comply with the voting population and the administrative boundaries. Indeed, quality issue is of first importance: there must not be any hole in the partition, polygons must not overlap and must be continuous, closed, without unnecessary arcs. These constraints make a good framework to target a good partitioning according to the content of the legal texts. Another way to assess the partition quality is to compare it to an existing reference such as the Cartelec map. Once obtained a set of contiguous polling station areas, the last step of the process would be to verify the complete consistence of the whole partition and to compare geometries between different data sources (Cartelec maps, city maps, etc.). This aspect of the work is not presented in this article, where we only show some encouraging results in automatic delineation of single polygons, regarding their geometry quality. Indeed, it allows to detect some errors or ambiguity in the spatial databases or even in the legal texts themselves. Despite a supposed rigorous process in writing the legal text to describe the polling station areas jointly used with accurate and detailed spatial databases, it is not so easy to find and build the polygon contours in a perfect way. These issues are discussed in the paper, which aims at:

- Proposing a method and an algorithm to retrieve polling station perimeters, including textual and geographical information;
- Assessing this method to find out errors in the spatial database or/and in the associated legal text;
- Highlighting and discussing some methodological issues to generate a perfect topological structure of polling station boundaries.

## 3 Geographical information retrieval

### 3.1 Finding regular expressions in text

Regular Expressions [17, 8] are powerful and current tools to search words in chains of characters [4, 3]. They are indeed often used for geographical purposes [12]. The term «Regex» somehow refers to a micro-language dedicated to find and replace generic patterns in strings. Most of the Regexp are developed within many languages, such as Perl, Python, Java, Racket and PHP, or using additional packages. The way to write relevant regular expressions is the main issue when dealing with regex. For instance, simple characters such as $*$ and ? can be used and combined with functions for quick string replacement or retrieval. Indeed, some patterns may become abstract in their structure. One good property of the regular expressions is their flexibility and their capacity to deal with uncertainty. On a theoretical point of view, they are somewhat finite-state machines, because they maintain an internal state number when they browse the chain of characters, without way back. Moreover, some extensions allow to recognize irregular patterns. We used the Regex package of Python related to QuantumGIS.

### 3.2 Research in geographical information retrieval

Geographical information retrieval is a large and interesting research domain [13], because it joins semantic and spatial issues [15]. Jones and Purves [11] identified a set of key problems related to geographical information search and retrieval, notably: ambiguity in reading place names and addresses [14], vague or fuzzy semantic of the words describing space, problems in textual and spatial indexation [5]. As we will see, these issues strongly impact our problem. In the field of spatial analysis, there are works on extraction of localized information [2] and improvement of spatial ontologies [16]. Other works propose to (re)build information from textual sources *a posteriori* (for instance, see [6] about travel stories). Making match geometrical and textual information is a geocoding process. The problem in matching spatial information is known to be difficult to solve, i.e. producing a perfect merging of both alphanumeric and geometrical data sources [18]. Here, the objective is to reduce the errors in partitioning as much as possible, in the sensitive context of universal suffering and with a spatial data quality purpose.

The method we propose is based on the successive recommendations from [12], especially in «geoparsing» :

- Looking for a direct matching of words in texts and spatial databases, due to the lack of information in the official voting rolls available in town councils;
- Using a system based on rules and in particular regular expressions adapted to texts having a certain regularity;
- Later on, extending the process to learning and approximation approaches in case of too much irregularity and if the two previous stages were inefficient (this is not developed in this article).

## 4 Handled data

### 4.1 Regular expressions applied to legal texts of polling stations

We used the package *RE* in Python, which allows to import sections of text from a corpus of polling stations.

Following examples are in French language. The sample starts by the following sentence: «Including the voters leaving in the area bounded by the street Ferruce even side...». «Rue» is street, «place» is square, «côté» is side, «pair» is even, «impair» is odd, etc.

*>>>sample_text="Comprenant les électeurs demeurant sur la partie du territoire délimitée par la rue Ferruce côté pair, la rue Puits de la Reille côté pair, la rue Balance côté pair, la place Puits des Boeufs côté pair, la place de l'Horloge côté pair,..."*

In this section we aim at finding out streets and squares using:

*>>>re.findall("(rue.*?|place.*?)côté",sample_text)*

The list obtained is then stored in the database:

*['rue Ferruce ', 'rue Puits de la Reille ', 'rue Balance ', 'place Puits des Boeufs ', 'place de l'Horloge ']*

We worked on the legal texts from Avignon. We can notice that there are recurrent words and expressions in those, according to a shared thesaurus.

Here is an example of a short legal text for a little polling station perimeter:

- Metadata about the text in the corpus:
  *Polling station label: BUREAU N° 101*
  *Polling station location: HOTEL DE VILLE - PLACE DE L'HORLOGE*
- Sequence of the sections:
  ***Comprenant les électeurs demeurant sur la partie du territoire délimitée par** la rue Ferruce **côté** pair, la rue Puits de la Reille **côté** pair, la rue Balance **côté** pair, la place Puits des Boeufs **côté** pair, la place de l'Horloge **côté** pair, la rue des Marchands **côté** impair, la place Carnot **côté** impair, la rue Armand de Pontmartin **côté** impair, la rue Sainte Catherine **côté** impair, la rue Lafare **côté** impair, la rue du Grand Paradis **côté** impair, la place Saint Joseph **côté** impair, la rue Palapharnerie **côté** impair du 15 à la fin, du quai de la Ligne à la porte du Rhône.*

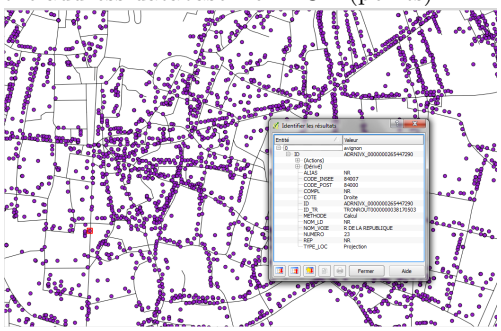
For the whole corpus of texts, some terms allow to find reliable instances with addresses and geometries: types of places (*street, square, dock, rail, gate*) or their names (*train station S.N.C.F*), or topological terms such as intersections of linear geometries (*crossroad*). Other words do not have an explicit location (for instance *the imaginary line joining* two crossroads). Also there are subtle differences in the ways similar texts are written with(out) parenthesis or space, which leads to potential mistakes: *côté pair*, *(côté pair)* or *( côté pair )* means the same. This type of problem is easily solved using regular expressions.

### 4.2 Geographical data

We mainly handle two types of complementary spatial data *i.e.* road networks and address locations (cf. Figure 2):

- The **map of roads from Navteq** includes the roads of Avignon. Attributes are names (street, square, docks, road, etc.) to locate the linear geometry. We tested the spatial database from Open Street Map, but its quality was not good enough to provide reliable results;
- The **address database from IGN** provides information about building numbers (points) and associated road name and side. It was useful to compare the results of geocoding to validate the extracted geometries from legal texts and Navteq spatial database.

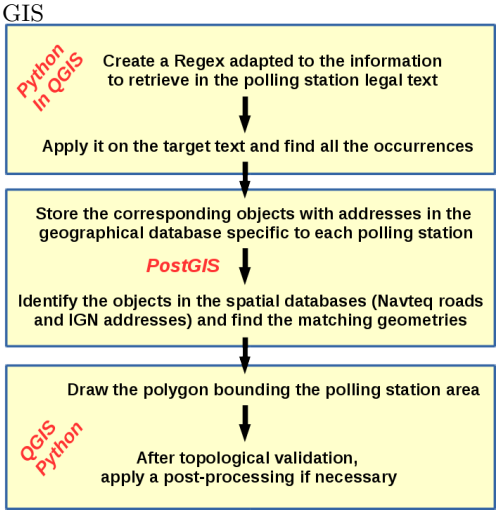
Figure 2: The Navteq network around Avignon, France with the address database from IGN (points).



## 5 Combining a GIS and an algorithm of information retrieval via Regex

We used Quantum Gis for viewing and analysing the polygonal geometries, interfaced with PostgreSQL/PostGIS, for structuring and storing the information, textual and geometrical as well (Figure 3).

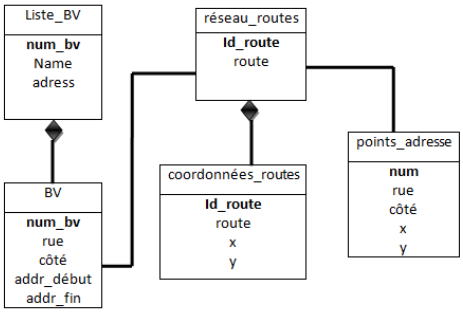
Figure 3: Algorithm of polling station area delineation with GIS




The *RE* package in Python provides several useful functions, such as *compile()* for compiling regular expressions to be executed or *findall()* to get the list of all the expression occurrences in the text.

The process, detailed in the Figure 3, consists in preparing a database in PostgreSQL/PostGIS to store the geographical data (roads, buildings), extracting the objects (places, streets...) from the legal texts, finding (in PostGIS), drawing (in QGIS) their geometry and eventually correcting geometry using a post-processing based on topology consistence.

Figure 4: Principal entities in the data conceptual model: polling stations, road networks and address points with their geometry




The Figure 4 shows the simple spatial database structure we built. Other intermediate and temporary entity classes are created on the fly (not detailed in this article).

## 6 Results: some polygons with various geometrical quality and topological consistence

Many tests were made on a set of polling stations. We provide in this paper a selection of typical cases to emphasize good results and also some limitations of the delineation process. As a text enumerates ordered (road sections (edges), the algorithm is in theory topologically consistent to close the polygon. However, some problems appear when looking at several polling stations.

The main errors in geometry result from dangle edges or duplicated geometries, because the complete road section referring to a name and an address can exceed the relevant crossroad. These cases are shown in the middle and on the right of the Figure 5 or in the Figure 6. These issues can be due to:

- A missing connection or road section in the legal text;
- A mistake in the spatial Navteq database (road name, for instance) that cannot be corrected by the IGN database;
- Existing little streets, trails or cul-de-sacs that keep the name of the principal street or road.

Sometimes, the geometry is incomplete: geometry misses a few road sections (edges of the graph) that make the polygon open. This case is also quite frequent and impairs the polygon topological integrity. This is illustrated in the left part of the Figure 5.

Figure 5: Main errors in polygon geometry and topology: holes, duplicated geometries, dangle edges

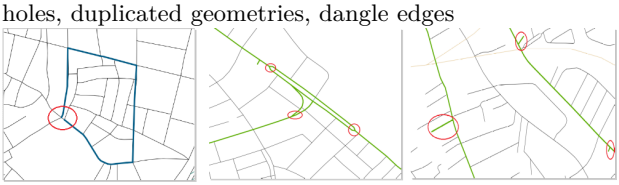


Figure 6: Basic geometries extracted for the polling stations 214 and 219 with long dangle geometries

A post-processing can solve most of the problems (cf. Figure 7). It consists in a verification of the topology integrity by imposing that:

- Two successive edges must be connected: this means that the algorithm fills the hole by the appropriate common section between the two edges; it becomes more difficult when there are too many missing data (for instance several missing edges that should be connected);
- On their both sides, all the edges must be connected to a unique other edge; this allows to drop dangle edges.

It is very interesting to notice the algorithm we developed was sometimes useful to verify (and to correct) some polygonal geometries provided by Cartelec in Avignon (cf Fig. 8). Although the first geometry obtained automatically (on left of the figure) was not perfect, at least it respects the main shape of the real footprint of the polling station area, which was split into two different polling stations indeed. It enabled us to check if the Cartelec map was not correctly updated for this polling station 219.

## 7 Conclusion and discussion

In this paper, we designed an algorithm that rebuild polling station area boundaries from legal text. It is especially applied in the French context of electoral partitions. Based on regular expressions, a GIS and a post-processing final stage, it is useful in most of the cases studied. However, it is subject to spatial data and/or legal text quality and needs improvement to be generalized. Indeed, the algorithm was only tested on a set of polling stations in Avignon. Its use needs to be enlarged to other polling stations from different towns and cities in France or even in other countries where similar problems may be faced by

Figure 7: Final correct, complete and accurate geometries extracted from legal text after post-processing (polling stations 214 and 216)


Figure 8: A polling station (nb. 219) drawn in Cartelec map (right) *versus* using the automatic algorithm (left).


the authorities. In France, a rapid glance on a few other legal texts shows that there may be significant differences in the ways the texts are written in different territories. This suggests that thesaurus should be locally adapted and that new learning approaches should be developed to perform the first stage of the algorithm based on regular expressions scanning texts. This may be the condition for improving the computation efficiency and flexibility and for generalizing its use as an operational tool for urban (at least French) authorities. Moreover, further work should be developed over a large series of polling station geometries, to test polygons correctness, to count outdated geometries and to assess misses, errors still remaining after applying the algorithm, by comparing the results with Cartelec maps and more accurate and updated polling station maps digitized by city offices. Also, the part of errors coming from either text or spatial databases (or both) should be estimated to infer a better knowledge of the process and to get reliable spatial partitions.

## References

[1] ACE. The ACE encyclopaedia: Boundary delimitation. Technical report, The Electoral knowledge network, www.aceproject.org, 2013. 224 pages.

[2] Frédérik Bilhaut, Franck Dumoncel, Patrice Enjalbert, and Nicolas Hernandez. Indexation sémantique et recherche d'information interactive. le moteur géosem. In *Proceedings of CORIA, Saint-Etienne, 28-30 mars 2007*, pages 65–76, 2007.

[3] Michael Fitzgerald. *Introducing Regular Expressions*. OReilly, 2012.

[4] Jeffrey E. F. Friedl. *Mastering Regular Expressions*. O'Reilly, 2006.

[5] M. Gaio, V.T. Nguyen, and C. Sallaberry. Typage de noms toponymiques à des fins d'indexation géographiques. *Revue Traitement Automatique des Langues*, 53 (2):143–176, 2012.

[6] Mauro Gaio, Christian Sallaberry, Patrick Etcheverry, Christophe Marquesuzaa, and Julien Lesbegueries. A global process to access documents' contents from a geographical point of view. *Journal of Visual Languages & Computing*, 19:3–23, 2008.

[7] Alain Garrigou. *Histoire sociale du suffrage universel en France, 1848-2000*. Paris, Seuil, 2002.

[8] Jan Goyvaerts and Steven Levithan. *Regular Expressions Cookbook*. O'Reilly, 2012.

[9] Olivier Ihl. Une ingéniérie politique. Augustin Cauchy et les élections du 23 avril 1848. *Genèses*, 49(4):4–28, 2002.

[10] Anne Jadot, Michel Bussi, Céline Colange, and Sylvano Freire-Diaz. Un outil d'analyse électorale en cours de création. CARTELEC, un SIG au niveau des bureaux de vote français. *Le monde des cartes. Revue du comité français de cartographie*, 205:81–98, 2010.

[11] Christopher B. Jones and Ross S. Purves. Geographical information retrieval. *International Journal of Geographical Information Systems*, 22(3):219–228, 2008.

[12] Jochen L. Leidner and Michael D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11, 2011.

[13] Christopher D. Manning, Prabhakar Raghava, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[14] B. Martins, I. Anastacio, and P. Calado. A machine learning approach for resolving place references in text. In *Lecture Notes in Geoinformation and Cartography*. Springer Verlag, 221-236.

[15] Christian Sallaberry. *Geographical Information Retrieval in Textual Corpora*. ISTE WILEY, 2013.

[16] Van Tien Nguyen, Mauro Gaio, and Christian Sallaberry. Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels. In *TIA'09 Toulouse 18-20 novembre 2009*, 2009.

[17] Andrew Watt. *Beginning Regular Expressions*. WROX, 2005.

[18] Antoine Widlocher, Eric Faurot, and Frédérik Bilhaut. Multimodal indexation of contrastive structures in geographical documents. In *Actes RIAO 2004, Avignon*, pages 550–570, 2004.

[19] Soumaya Yahiaoui, Didier Josselin, Christèle Marchand-Lagier, and Johnny Douvinet. Vérification et (re)construction automatiques des limites des bureaux de vote par l'étude des textes juridiques. In *SAGEO 2014*, Grenoble, France, November 2014.

[20] Soumaya Yahiaoui, Didier Josselin, Eric San Juan, and Christèle Marchand-Lagier. Délimitation géométrique semi-automatique des bureaux de vote à partir de textes juridiques et d'information géographique numérique : enjeux et difficultés. In *Congrès AFSP 2015*, page 14 pages, Aix en Provence, France, June 2015.