CO-408

HOW TO GEOLOCALIZE AN ADMINISTRATIVE FILE BASING ON ITS ADDRESSES DATA ? A PROPOSAL TO ACHIEVE A BETTER GEOCODING.

CHARIF O., SCHNEIDER M., OMRANI H.
CEPS/INSTEAD, DIFFERDANGE, LUXEMBOURG

For many years, researchers have presented the geocoding of postal addresses as a challenge. Several research works have been devoted to achieve the geocoding process. This paper presents theoretical and technical aspects for geolocalization, geocoding, and record linkage. It shows possibilities and limitations of existing methods and commercial software identifying areas for further research. In particular, we present a methodology and a computing tool allowing the correction and the geo-coding of mailing addresses. The paper presents two main steps of the methodology. The first preliminary step is addresses correction (addresses matching), while the second caries geocoding of identified addresses. Additionally, we present some results from the processing of real data sets, and an example application of the result produced by this tool. Finally, in the discussion, areas for further research are identified.

KEYWORDS
address data, data matching, geocoding

1. INTRODUCTION

A lot of research works have been devoted to geocoding of postal addresses. The interest in this topic is supported by the need to transform postal addresses into geographical coordinates which are essential for various domains of scientific and social research. The benefits of the address geocoding precision are numerous. Geocoding can be used for a wide range of applications such as market segmentation, demographics, geo-spatial distribution of plants, sales territories, taxes, elections. Geocoding is also a very important tool to target certain demographics characteristic. The results of geo-coding have provided fundamental components for wide variety of research works in many fields (e.g. health [4], crime analysis [8], political science [5], computer science [7], etc.). The geocoding operation plays for example an important role for marketing in companies; it helps to cluster peoples with specific characteristics that might be interested in their products.

Many research centre and companies have developed free and commercial geocoder. A big number of these softwares use the linear interpolation method to calculate the spatial coordinates. This method estimates the coordinates of an address using the coordinates of bordering addresses of the street where the address is located. Many research papers [1], [7] have described the error in localization produced by using the linear interpolation method. It was mentioned in [1] that the error in localization can reach 3 kilometres (the distance between the true position and the estimated localization). In addition to the localization error, a big number of the developed tools do not take into account misspelled and abbreviation errors which are made while writing the postal addresses.

These tools are not able to deal with miswritten addresses such as miswritten road name, city name, etc. After studying some of the existing solutions for geocoding, we decided to develop our own geocoder. The developed tool is able to detect and correct errors as well as to deliver the better precision in term of the localization of the input addresses. The structure of the paper is as follows. First, we present a brief overview about geocoding. Second, we describe the developed methods. Subsequently, the Results of processing administrative files are summarised, and an application of the result is presented. Finally, we conclude the paper and show some areas for future development.

2. OVERVIEW

Most existing works in the field of geocoding are developed based on the structure shown in Fig. 1. The geocoding process is divided into three main steps:

1. Structuring and normalizing: it consists of cleanong and normalizing the input address.
2. Record linkage: it allows finding a match of the inputted address in the reference database.
3. Geocoding: it calculates coordinates of the indentified address.

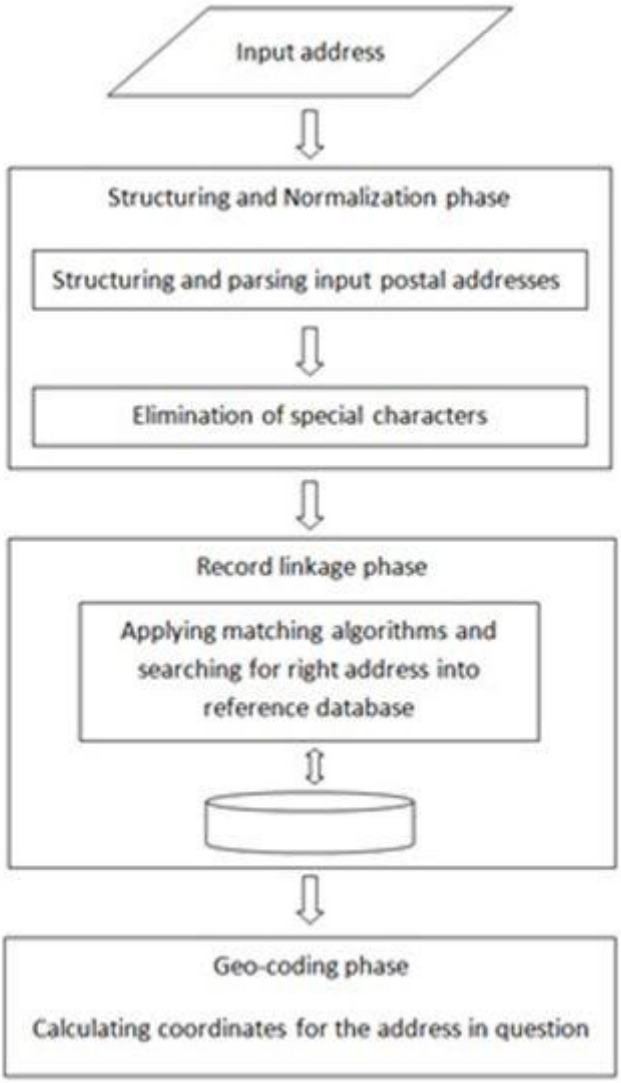


Figure 1 : General structure for geocoding process

Existing research works usually differ with respect to the methods which are used on each step of the geocoding process. Fig. 2 summarizes methods currently available for each step.

• Structuring and normalizing step: this step is required for cleaning and structuring the input address. The most difficult part of this step is the normalization where each different part of a postal address (postal code, address, road name, etc.) must be identified from a complete input address. Fig. 2 presents details about different methods already used in this step.

• Record linkage phase: It allows comparing names and address information across to pairs of data sets (the reference data set and the input data set) to find out if they are describing the same entity. It is during this step that errors in writing an address will be detected (methods are shown in fig 2)

• Geocoding: the final step of the process is to calculate the spatial coordinates. This step finds the coordinates while considering the desired scale (see methods shown in fig.2)

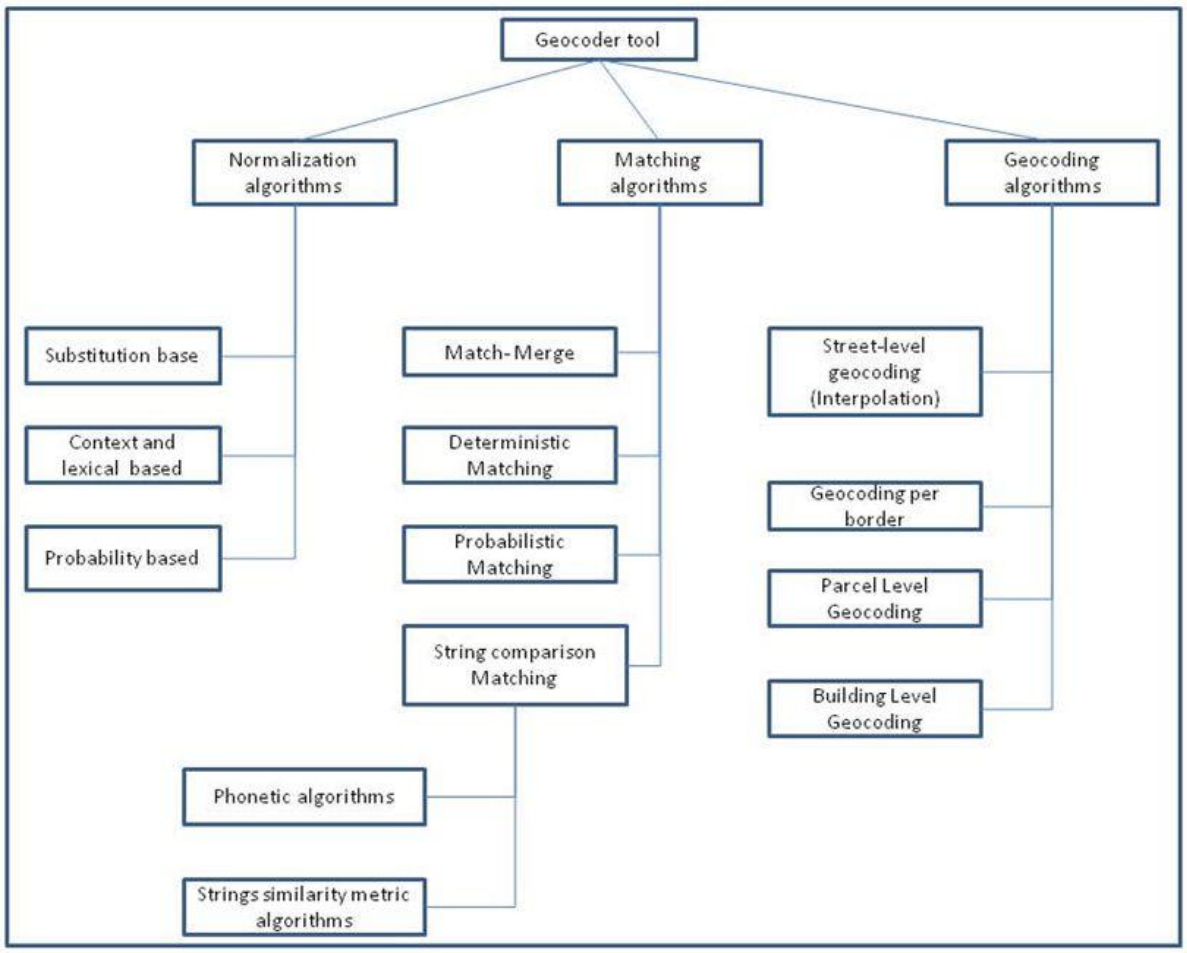


Figure 2: Methods used in geocoding process

3. METHODOLOGY

We have developed a general purpose tool for geocoding while taking into account the particularities of our case of study (Luxembourg). Consequently, some parts of our methodological choice have been influenced by both characteristics of Luxembourgish postal address system and the type of files that we anticipated to process. Below we present and the justify our methodological choices:

3.1 Step 1: Structuring and normalizing

This tool has been developed to process administrative files, where the postal addresses were divided into fields (road name, postal code, municipality, etc.). Thus, we didn't need to perform a complicated normalization technique that parse the input address into fields. In some cases, we used substitution based normalization in order to distinguish two parts of an address that were coded under the same variable (e.g. L-3123 where the letter "L" represents country Luxembourg and 3123 stands for postal code). This method relies on the type of the fields to identify them (e.g. postal code is usually a number and country is a string). It divides the input address into tokens by using "space", "comma", etc as a separator. It will then associate tokens to fields that have the same type. On the other hand, the fact that Luxembourg is a multi language country (i.e. Luxembourgish, French and German) has brought up the need of cleaning (eliminating special characters) and standardization step.

3.2 Step 2: Record linkage

Besides geocoding, we also developed a tool able to correct mistakes produced while inputting data. The decision in this step was very important for the success of the work presented in this paper. Thus, the biggest work lies in the effort to find an algorithm able to detect and correct mistakes while matching the input address with addresses in the reference database. Although the first two choices (Match-Merge and deterministic, see fig.2) were very simple to implement, they were not able to deal with complicated misspelled and mistakes. According to Dey [3] the string comparison methods have shown higher reliability than probabilistic methods.

Following the results presented in table 1 which present result obtained by applying different String similarity metric methods to one road name written in two different way ("AVENUE J.F.KENNEDY" and "AVENUE J-F KENNEDY"), we have noticed that the "Jaro", "Jaro winkler", "Levenshtein", "Mongo Elkan" and "soundex" were the best in detecting misspelling errors.

	Jaro	Jaro Winkler	Levenshtein	Mongo Elkan	QGrams	Jaccard	Soundex
Similarity index	0.925	0.970	0.888	1.0	0.75	0.25	1
Processing time	0.013	0.014	0.058	0.86	0.043	0.0008	0.020

Table 1: Similarity calculation results for matching "AVENUE J.F.KENNEDY" and "AVENUE J-F KENNEDY"

	Jaro	Jaro Winkler	Levenshtein	Mongo Elkan	QGrams	Jaccard	Soundex
Similarity index	0.893	0.957	0.75	0.888	0.571	0.5	1
Processing time	0.009	0.010	0.043	1.238	0.032	0.001	0.017

Table 2: Similarity calculation results for matching "RUEDESARDENNNES" and "RUEDESJARDINS"

On the other hand, the results (shown in table 2) of comparing two roads with very similar name demonstrate that "Levenshtein" method is more reliable than "Jaro", "Jaro Winkler", "Mongo Elkan" and "Soundex". Thus we decided to combine two techniques of string comparison "Livenshetein distance" [6] and "vectorial"approach (e.g. Q_Grams algorithm [2]). The "Levenshtein distance" calculates the number of operations (i.e. add, remove, substitution) which is needed for passing from one string to another, which helps to detect and correct the misspelled errors. Yet this method is not able to detect abbreviation based errors. This type of errors requires the intervention of the "vectorial technique" which consists of dividing the compared names into tokens or words. Fractions of each matched name will then be compared (by comparing the two words using "Levenshtein distance" or just by comparing first letters of the two words) while a percentage of similarity is calculated. These choices were made by considering the processing time and the reliability of the similarity metric results (tables 1 and 2).

The matching procedure begins with verifying the existence of the couple postal code and road name by querying the reference database. If the answer of this query is null we always assume that the postal code is correct. The reason for this is because errors are most likely to be committed while inputting text data. We then create a list of roads which are associated with the input postal code. An algorithm is then executed to match the input address with captured road list. If the matching did not succeed then the same procedure is repeated but with a list of roads associated to the input Municipality. In order to accelerate the processing time, we have created a knowledge database which helps to memorize the variants of names writing (errors already detected). This knowledge database becomes richer as we run a new file process.

3.3 Step 3: Geocoding

It has been shown in [9] that the quality of geocoding has a big influence on the results of analyses which use it. According to [1], the error in localization which is produced using parcel geocoding method is significantly smaller than the error in localization which is produced by using street geocoding method. These two facts and availability of a database which contains coordinates for buildings have encouraged us to use the building localization method. In case the reference data base does not contain the input building, two solutions have been adapted. First, in case of the existence of a building on the same of the missing address we try to predict the localization of the missing address using a combination of building geocoding and kind of linear interpolation. we called this solution "geocoding by nearest neighbor". we have developed 2 methods (parametric model and least square mean curve fit).

Parametric model:

This method is based on the conversion of a line defined by two point A(xa,ya) and B(xb,yb) into a parametric vectorial equation which facilitate the way to find a point along a line once an the distance (number of step) between this point and any other point on the line is available. The algorithm of this method is as followed:

1. Calculate the slope given by: $m = \frac{y_a - y_b}{x_{a-x_b}}$, for each steps one of x, y change m unit. the vector of variation is <1,m>
2. Calculate the number of steps that separate the two addresses and the direction of movement (backward or forward):
   numStep $= \frac{(n-numNN)}{2}$; where n is the missing building number and numNN is the nearest number to n.
3. calculate the mean distance given by:
   $$mean\_distance = \frac{\sqrt{(x_b - x_e)^2 + (y_b - y_e)^2}}{\frac{n_b - n_e}{2}}$$
   where x_b, y_b and x_e, y_e are respectively the coordinates of the address with the smallest (n_b) and the biggest (n_e) building number in the road in which the missing building exist (with the same parity (side)).
4. calculate the coordinate of the missing building using the following equation:
   $$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} x_{numNN} \\ y_{numNN} \end{pmatrix} + \begin{pmatrix} mean\_distance * |cos(slope)| \\ mean\_distance * |sin(slope)| * m \end{pmatrix}$$

Least square mean curve fit:

This method is based on the idea of finding the curve fit of a line given a set of m points on it. It consists of constructing a polynomial $p(x)$ of degree m-1 that minimize the square error given by:

$$E = \sum_{i=0}^{m} [y_i - p(x_i)]^2$$

In figure 3, we pretend that the building number 27 is missing and we have tried to predict its localization using the two developed methods. The figure 3 shows that least square mean curve fit give a better results. This method is very in case of the existence of curves in the road in question. On the other hand, this method shows lack of precision when trying to predict the localization of an address that exists in a straight road where one of the coordinates is almost unchanged along the road.

Figure 3 : Two methods to geocoding by nearest-neighbour and errors

second, in case of the non existence of an building on the same side of the missing address, we associate to this address the coordinates of road barycentre in whish it exists.

4. RESULTS

We present in table 3 the result of processing of four data sets from different administrative sources. The first three are results of geocoding data sets containing addresses of six test municipalities in Luxembourg. The fourth is the result of processing a data set that contains addresses from all over Luxembourg. The developped tool contains interactive, user friendly interfaces which facilitate the setup of settings needed for data sets processing as shown in figs 2 and 3.

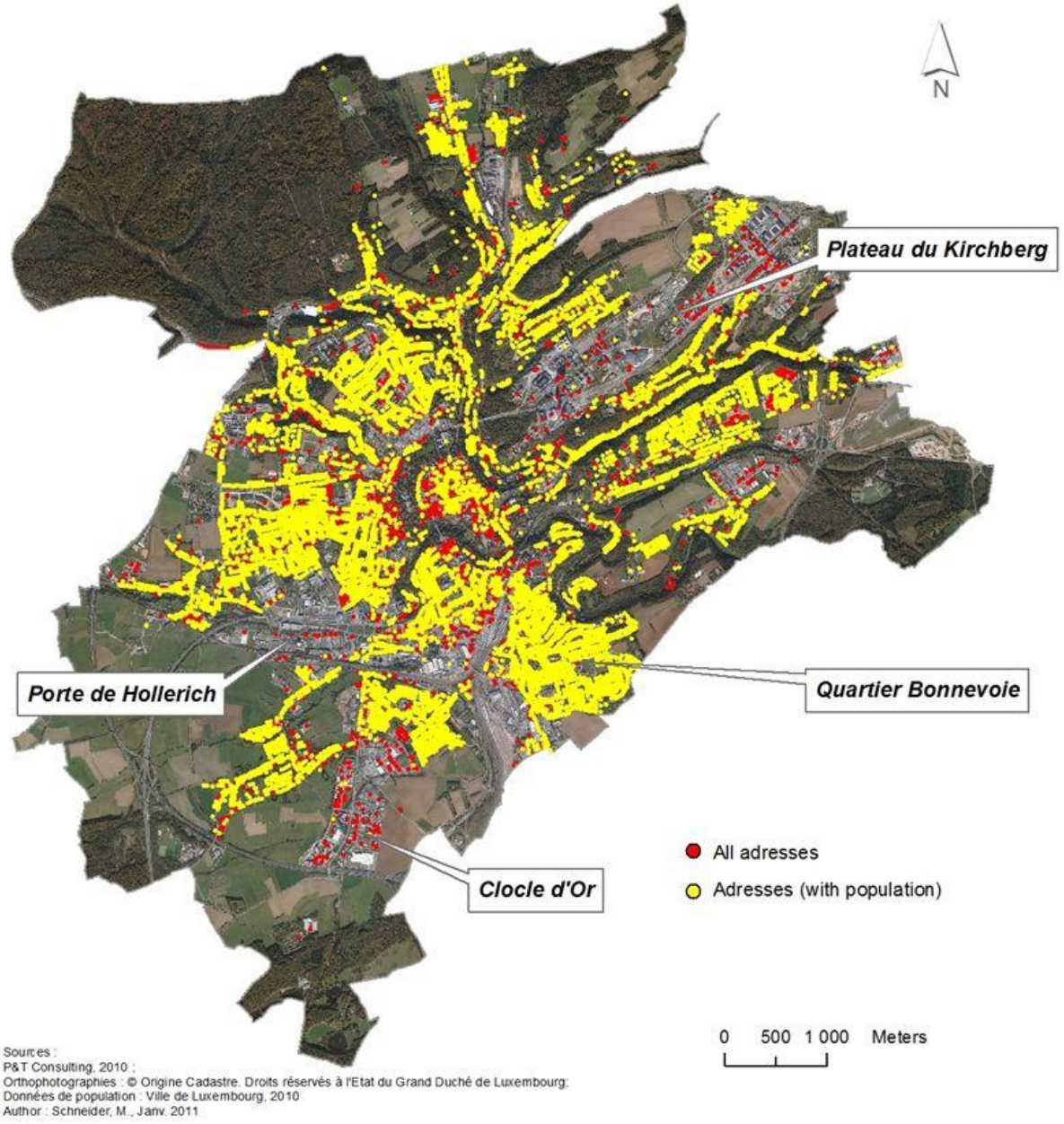
	Total record	Geocoding percentage	Missing data percentage	Building geocoding percentage	Geocoding nearest Neighbor percentage	Geocoding by road barycentre
Data set 1	19409	98.20%	0.015%	88.83%	10.4%	0.45%
Data set 2	35594	97.51%	0.073%	86.32%	12.31%	1.37%
Data set 3	38672	81.98%	16.97%	84.69%	13.43%	1.88%
Data set 4	457339	95.43%	0.024%	97.87%	1.68%	0.45%

Table 3: Results of processing four data sets from different administrative source

5.APPLICATION

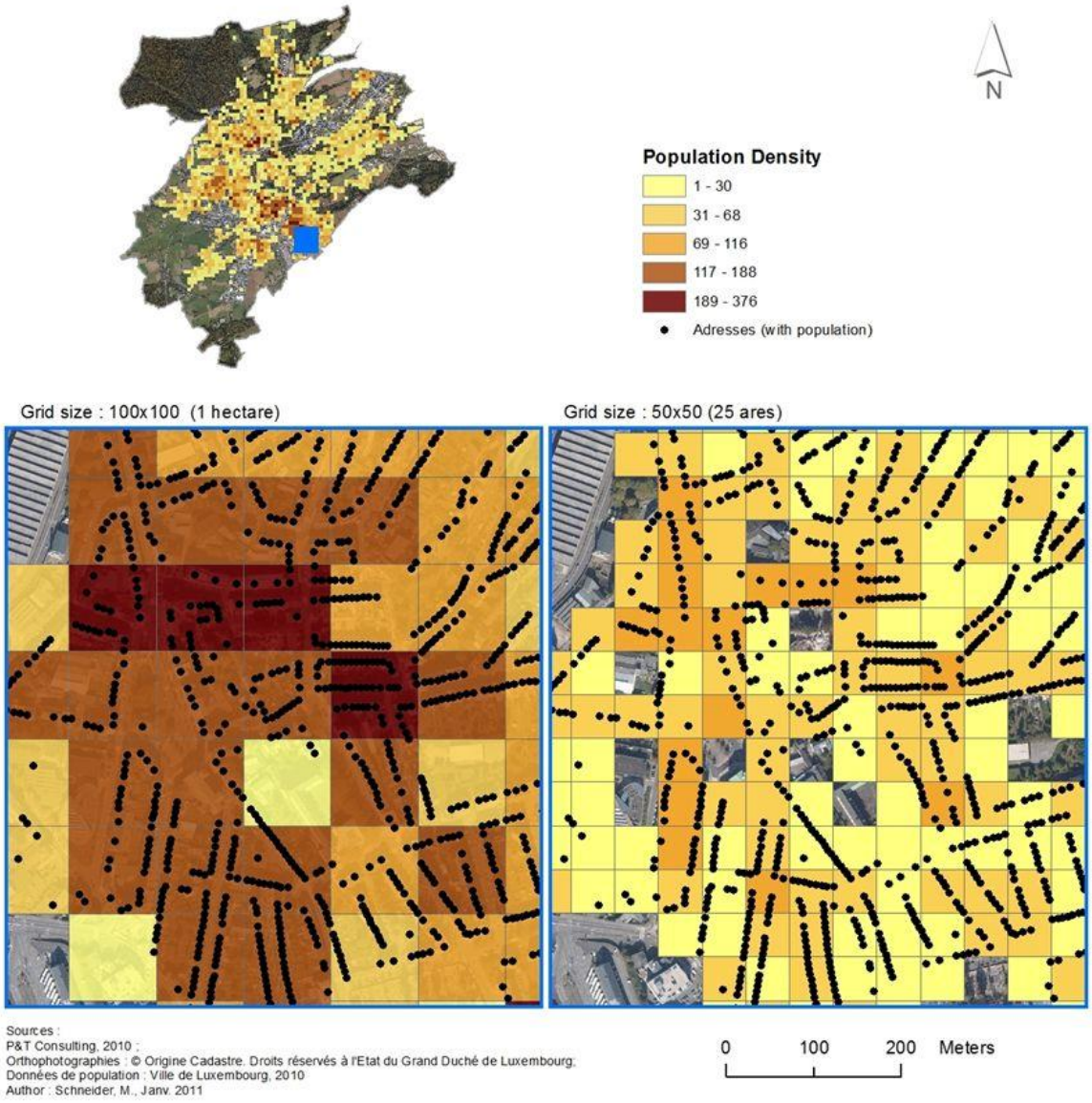
The Population file of Luxembourg City allows for instance to associate a person to an address thus to identify the housing zones in an area. The file of addresses counts more than 18700 mailing addresses. A matching of this file with the population register (> 80000 individuals) shows those 15200 addresses is associated to at least one person. Thus more than 81 % of the mailing addresses in the city are identified as residential places (see Map 1). One of the interests of this kind of mapping is the identification of non-residential areas e.g." The Kirchberg plateau" in which mainly European institutions and firms are localized. The map 1 shows clearly that the main street (Boulevard J. F. Kenndy) of the "The Plateau Kichberg" is devoid of residential buildings . The same situation can be encountered in the south side of Luxembourg city ("Cloche D'or" employment hub, "Porte de Hollerich", etc.) where a remarkable lack of

residential buildings can be noticed. On the other hand, the "Bonnevoie" quarter in the east side of Luxembourg shows an opposite situation of the two previous examples where almost all constructed space is devoted for residential buildings.



Map 1: Addresses of Luxembourg City

This type of simple maps presents a primary approach that can be used by urban planners-developers to answer questions related to multifunctional urban area (e.g. residential and commercial), land consumption,... . The tool we developed process administrative files provided by various sources by making a correction to the mailing addresses (if needed) and then geocoding these verified addresses. This procedure delivers the information at the finest scale (mailing address) which can be spatially aggregated to simplify the analyze of the produced maps using these information. As a simple example, in map 2, we have used mailing addresses to show the population density variation in a squared grid of 100x100 and 50x50. The area studied in this map is a small part of the "Bonnevoie" quarter (south-east of Luxembourg City). The map 2 shows that once you have the data at the finest scale you can easily change to a higher scale using simple manipulation in many tools that implement a geographic information system.



Map 2 : Density of population (aggregation)

5. CONCLUSION

In this paper, we have presented two new methods. The first one is for record linkage and the second is for coding. These two methods have given good results with more than 95% percentage of success. We have implemented and developed the computation tool using Java programming language (see figure 4,5,6). However in the future, a normalization of the input address module must be added to this tool.

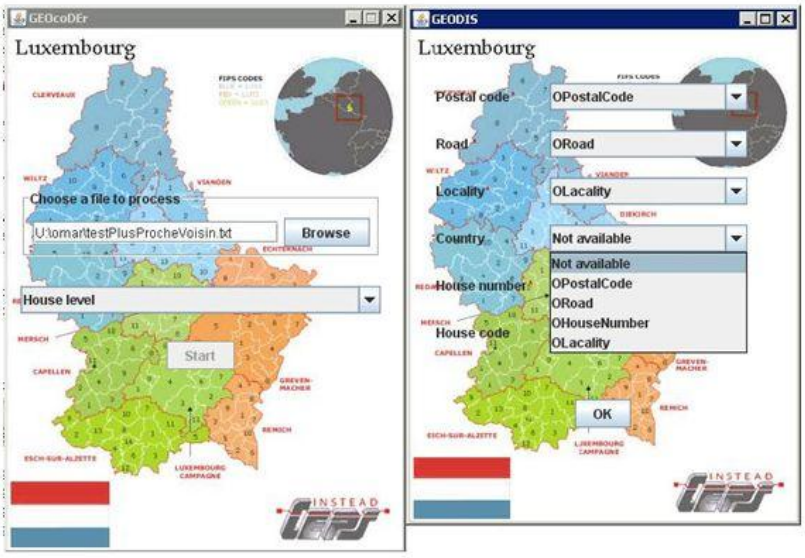
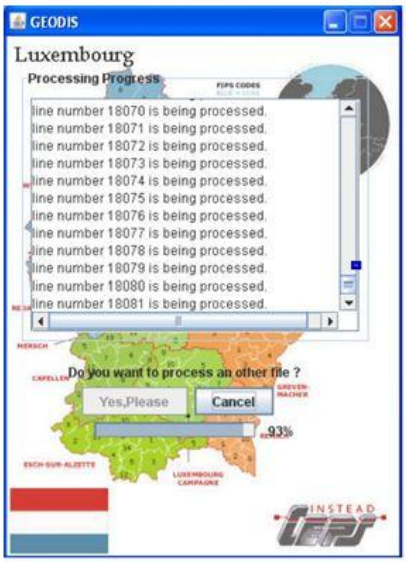


Figure 4: Select file and setup processing settings



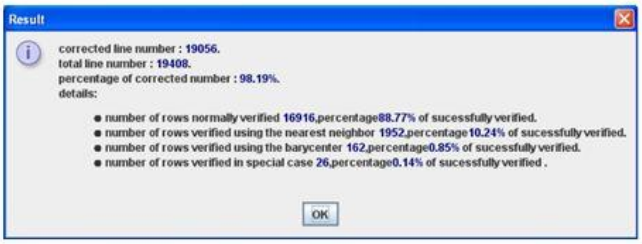


Figure 6: Result user interface

Figure 5: Processing progress user interface

REFERENCES

[1] M.R. Cayo and T.O. Talbot. Positional error in automated geocoding of residential addresses. International journal of health geographics, 2(1):10, 2003.

[2] T. Churches and P. Christen. Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4(1):9, 2004.

[3] D. Dey et al. A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases. IEEE Transactionson Knowledge and DataEngineering, 14(3):567 - 582, 2002.

[4] G. Rushton et al. Geocoding in cancer research : a review. American Journal of Preventive Medicine, 30(2):16-24, 2006.

[5] M. Haspel and H.G. Knotts. Location, Location, Location: Precinct Placement and the Costs of Voting. The journal of politics, Cambridge University Press, 67:560-573, 2005.

[6] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 10:707, 1966.

[7] R. Bakshi et al. Exploiting online sources to accurately geocode addresses. ACM-Gis, pages 194–203, 2004.

[8] J.H. Ratcliffe. Geocoding crime and a first estimate of a minimum acceptable hit rate. International Journal of Geographical information Science, 18(1):61–72, 2004.
[9] P.A. Zandbergen. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health, 7(37):1-13, 2006.