AUTOMATIC DERIVATION AND CLASSIFICATION OF HOUSES ON A CADASTRAL MAP

Maureen Rengelink, Peter van Oosterom, Wilko Quak and Edward Verbree

Delft University of Technology, Department of Geodesy
The Netherlands

ABSTRACT

In this paper we describe a project for the automatic derivation and classification of houses based on the Cadastral map and Cadastral administrative information. The Cadastral map contains the boundaries of buildings, which sometimes can be subdivided into smaller living units by using ownership parcel boundaries; e.g. in case of a row of houses. This method will not work in case of a row of rental houses, because they are all owned by the same legal entity and the Cadastral map does not define boundaries between the individual living units. In this case an additional data set, Address Coordinates Netherlands (ACN), helps identifying these living units.
A house will be assigned one of six different classes. This paper describes a method to derive and classify the houses and also evaluates the obtained results.

1. INTRODUCTION

In the Netherlands there is a growing need for a map product, on which every house has been assigned a specific type, such as 'apartment building', 'free standing house', and so on. One of the purposes of such a product is the analysis of trends in house prices in the country, which depend on the house type. In this paper a house is defined as a single unit used for one family to live in. Currently, there is no digital map of the Netherlands, on which individual houses are modelled as objects. However, a spaghetti map of the boundaries of buildings is available. These boundary lines must be combined into polygons and split up into individual living units. We plan to derive the requested map from spatial and thematic data at the Dutch Cadastre, in three steps. First we make a polygon layer of the boundaries of buildings, so that each building appears as a separate polygon. Second, a map overlay of the buildings and parcels is performed to generate an integrated map of parcels and buildings. In the third step, this map together with administrative Cadastral information will be used to perform the automatic classification based on decision-rules. The goal of this project is to find out whether it is possible to automatically derive and classify the houses on the cadastral map, and whether the quality of the automatic classification is satisfactory.

The remainder of this paper is structured as follows. In section 2 an overview of the classes into which the houses must be classified is given. Section 3 describes the datasets that are used for the classification. Section 4 describes the classification process itself. Section 5 gives an analysis of the quality of the classifications of section 4. The current status of the project, future plans, and conclusions can finally be found in section 6.

2. DIFFERENT HOUSE CLASSES

In the classifications process the following classes are assigned to the buildings:

- **middle house in a row:** A house in the middle of a row of terraced houses. This type will be given the code “M”.
- **end house in a row:** A house at the end of a row of terraced houses. This type will be given the code “E”.
- **free-standing house:** A house standing on a parcel without any other houses. This type will be given the code “V”.
- **two-under-a-roof:** This class of house is very common in the Netherlands; it consists of one building that consists of two houses standing on their own ground. This type will be given the code “T”.
- **apartment:** In the administrative cadastral database parcels related to apartment buildings are explicitly marked. This type will be given the code “A”.
- **special:** A house of a type not described above. This type will be given the code “S”.

In addition to this classification, all houses are also assigned the predicate "rented" or "bought". The approach for classifying the rented houses will be different from the approach for identifying the privately owned houses. This is because the Cadastre in the Netherlands only registers ownership of cadastral objects and not the rental of living units. For instance, there will be one building (representing a row of houses or living units) on one large parcel when one organisation owns them.

3. DATASETS USED

This section describes the different datasets used for the classification. All datasets are maintained by the Cadastre. From these databases we use the following data for the classification (and detection) of the houses:

The buildings on the Cadastral map

The buildings are stored as line elements (polylines and circular arcs) and text labels are used for house numbers and street names on the Cadastral map. The current map does not represent the buildings as area features (explicitly). All lines have an attribute code, which indicates the type of line. With this attribute the proper *boundaries* of main buildings (excluding the boundaries of barns and annexes) can be selected and transferred into a separate layer with only the boundaries of all main buildings.

The Cadastral parcels and administration

The Dutch Cadastre registers information in AKR and LKI. AKR stands in Dutch for “Administrative Cadastre Registration”, and contains all administrative information. LKI stands again in Dutch for “Landmeetkundig Kartografische Informatie” and contains all

geometric information. It contains the shapes and location of all cadastral parcels and other information to make a cadastral map (figure 1). For every parcel in LKI, there are many thematic or legal attributes in AKR:

- **Ownership Information:** Every parcel is owned by one or more subject. A subject can be either 'natural' (a person), or 'non-natural' (organisations etc.). The address, where an owner can be reached, is also included in this information (subject address).
- **Object address:** A parcel has one or more (e.g. several rental houses on one parcel) addresses. If the address of the object is different from the address of the owner, this is an indication that the house is rented.
- **Apartment Indication:** It is explicitly marked if a parcel is part of an apartment complex.
- **Culture code:** Every parcel has a two-digit landuse code, which gives an indication for the main land use. This code can be used to separate houses from other types of buildings.

ACN

The ACN (Address Coordinates Netherlands) [4] is a database with the geographic coordinates of all the addresses in the Netherlands recognised by the PTT (Royal Dutch Mail). The geographic position of the Address Coordinate inside a parcel dependents on the actual situation:

- If there is only one address connected to a parcel, the address co-ordinate can be anywhere in the parcel, not necessarily inside the building.
- If there is more than one building on a parcel, the address co-ordinate will be placed inside the building to which the address belongs.
- If a building has more than one address (for example apartment buildings), the address coordinates will be placed inside the building. Sometimes all addresses have the same coordinates inside the building; sometimes different addresses have different coordinates.

The ACN data plays a key role in the classification of rented houses; see Sections 4 and 5.

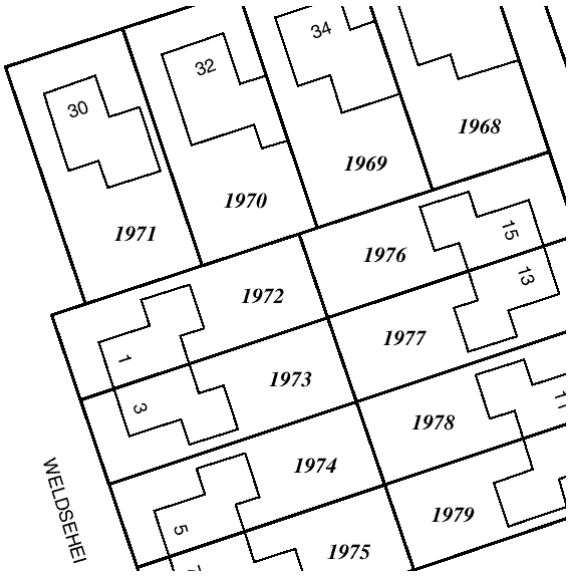

Figure 1: Part of Cadastral Map.

4. THE PROCESS OF DERIVATION AND CLASSIFICATION

This section describes the process of the derivation and classification of the buildings. It will take place in three phases: In the *first* phase, we extract and pre-process the information from the database to build two map layers, one containing all buildings and the other containing all parcels. In the *second* phase, the two layers are overlaid and a new map layer is created which contains both outer (original building boundaries) and inner (derived from parcel boundaries) walls. The areas defined by these walls can represent a part of a building (living unit) or an entire building (in case there are no inner walls). These areas will be called house Covered Parts of Parcels (CPPs). So, the original buildings can be split into several pieces using the boundaries of the parcels to find the inner walls. See figure 2 for an example of an overlay and the resulting house CPPs. The *third* phase automatically classifies all house CPPs using the derived geometric, topologic and related administrative information.

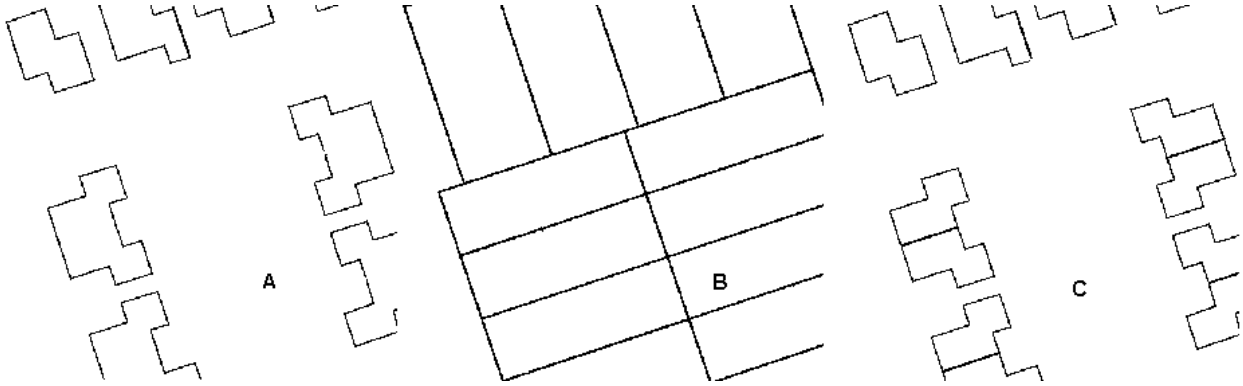

Figure 2: The overlay performed. "A" are the buildings, "B" are the parcels, "C" is the result, which contains the house covered part of parcel (CPP).

Transforming building boundaries into house Covered Parts of Parcels (CPP)

After extracting all buildings boundaries and parcel from the database, the second phase performs a map overlay. For this step we are using two alternative approaches: one is to use the Computational Geometry library CGAL [2] and the other one is to use ArcInfo [5].
In order to use the CGAL C++ library we extract the layer data from the database and read it into the overlay software. Before the overlay is performed some pre-processing should be done on the data. For example, adjacent parcels with the same owner should be merged. The resulting overlay, together with the administrative data will be copied to a geo relational database (such as Ingres [1]). Parallel to CGAL we have also used ArcInfo to perform the data clean-up and also the overlay. In the ArcInfo package the following steps were performed:

1. In the first step all outer walls of main buildings are selected. Internal walls are also sometimes available in the Cadastral map, however these are not helpful for the detection of buildings and are not considered.
2. Snapping dangling edges. An outer wall of a building should be a topologically closed loop. If we have a boundary which is not part of a loop (a dangling edge), one can try to snap such an endpoint to a nearby boundary to form a closed loop. This is done with the ArcInfo command 'snap'.

3. A polygon coverage of the lines is created by the ArcInfo ‘clean’ command. This operation takes care of intersecting building boundaries.
4. The result is a polygon layer with buildings represented as area features. Although there should be no internal walls in this layer, it appears there are a few of them. This is because sometimes internal walls are labelled as external walls in the Cadastral map. These walls are removed by deleting all lines that have a building to both sides.
5. All buildings are given a unique number.
6. The parcels already form a topologically correct map layer, so no pre-processing is needed.
7. An overlay between the parcels and the buildings layer is performed. This results in a map layer where buildings are split up when they are standing on different parcels.
8. Some polygons in the resulting overlay are very small; these (sliver polygons) are removed with the ArcInfo ‘eliminate’ command. Currently all polygons less than 20 square meters are removed.

In this resulting integrated map layer we will be able to pose the following queries: Is this building located completely within one parcel? How many mail-addresses are connected to this building? In the database a table CPP (house Covered Part of Parcel) has been created with the following attributes:

```
CPP(unique_id, shape, parcel_id, num_of_neighbours, area, area_entire_build)
```

Classification of the house CPPs

In the third phase of the process we are going to use the above-described database to classify all the houses. Writing rules for the different house types and expressing these, as SQL queries on the database will do this. An example of such a rule might be: A building on two parcels with different owners, is a 'two-under-a-roof'. By applying these rules to the database, the houses will be classified. Figure 1, the house on parcel 1971 is standing on one parcel and will be classified as a ‘free standing’ house. The building on the parcels 1972 and 1973 is standing on two different parcels. This will be classified as two houses of the type 'two-under-a-roof'. Other classification rules that can be applied are: A building with one owner with many addresses, probably is a building with rented apartments. Also the area of a building can be used to find a correct classification. First we link the administrative information (from AKR) to the house CPPs. This join is performed using the unique parcel-identifier.

```
CPP(unique_id, shape, parcel_id, num_of_neigbours, area, area_entire_build,
Appartmentcode, Objectaddr, subjectaddr, natural_person_code, culture_code,
area_parcel)
```

Next several new attributes must be calculated from geometric relations and the data from ACN: Number of buildings on a parcel, number of address co-ordinates inside house CPP, number of address co-ordinates inside parcel.
Final step now is to add two more attributes to our table and to use the information we have collected so far to fill these two columns. The two columns are "sector" and "type", which could be translated as house-sector and house-type. House-sector will contain the information whether it is a rented (code “H”) or owned (code “K”) by the people who live in it. And housetype will contain the assignment to one of the 6 classes given in Section 2.

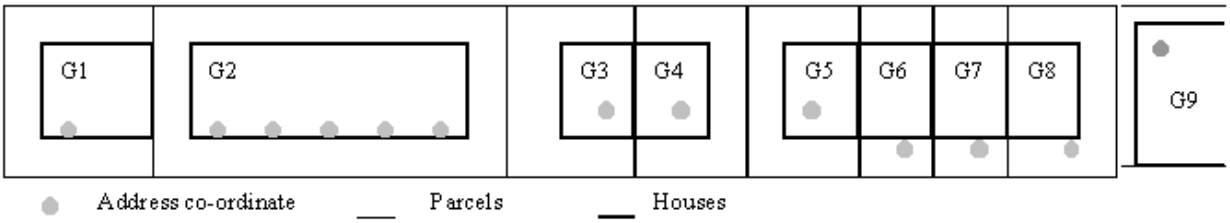

Figure 3: Graphical representation of situation described in table 1.

Unique_id	shape	parcel_id	Number of neighbours	area	Area -entire building	Appartmentcode	Objectaddr	subjectaddr	Natural_person_code	Culture_code	area_parcel	# parcels_under_building	# buildings_on_a_parcel	# address co-ordinates_in a CPP	# address_co-ordinates_inside_parcel	sector	Type
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16		
G1	Poly	Id	0	150	100	G	1234AA0003-	1234AA0003-	N	11	200	1	1	1	1	K	V
G2	Poly	Id	0	450	450	G	1234AA0005-	9999ZZ0111-	O	11	600	1	1	5	5	H	S
G3	Poly	Id	1	100	200	G	1234AA0015A	1234AA0015A	N	11	150	2	1	1	1	K	T
G4	Poly	Id	1	100	200	G	1234AA0015B	1234AA0015B	N	11	150	2	1	1	1	K	T
G5	Poly	Id	1	100	400	G	1234AA0017-	1234AA0017-	N	11	150	4	1	1	1	K	E
G6	Poly	Id	2	100	400	G	1234AA0019-	1234AA0019-	N	11	120	4	1	0	1	K	M
G7	Poly	Id	2	100	400	G	1234AA0021-	1234AA0021-	N	11	120	4	1	0	1	K	M
G8	Poly	Id	1	100	400	G	1234AA0023-	1234AA0023-	N	11	150	4	1	0	1	K	E
G9	Poly	Id	0	800	800	A	1234BB0001-	9999ZZ0111-	O	12	999	1	1	24	24	H	A

Table 1: Example of table house CPP.

In Figure 3 a few example situations are drawn. Table 1 corresponds to those examples are given. Remarks about the table:

- Q15 and Q16 are calculated by finding for each address co-ordinate the house and parcel it falls in with the help of a point-in-polygon-test, each time a house or parcel is found its number is added the number of addresses belonging to the address co-ordinate.
- The addresses of both the object and the owner (=subject) consists of the postcode (4 numbers and two letters), the house number and sometimes an additional letter.

To give some idea about our determination rules, some of them will be discussed here. If the number of neighbours (Q4) is zero and the owner is a natural person (Q10) and there are none or only one address co-ordinates (Q15 and Q16) in both the house CPP and its parcel and the owner lives in this house it's a **free-standing** (V) house. With the use of the qualifiers this rule could look like this:

```
IF ((Q4=0) AND (Q10="N") AND (Q15<2) AND (Q16 <2)) then type = V
```

If the number of neighbours is two (Q4) and the number of parcels under the original building

is more than 2 (Q13) and there is one address co-ordinate on the parcel (Q16) the house should get the classification **middle-of-a-row**. With the use of the qualifiers this rule could look like this:

```
IF ((Q4=2) AND (Q10="N") AND (Q16 = 1) AND (Q13 > 2)) then type = M
```

The end of a row is almost the same, except that Q4 than has to be equal to 1. To find whether a house is rented or not the owner has to live at the house address, so comparing Q8 and Q9 could be an indication. But to make this classification hold stronger also looking at whether the owner is a natural person or not is better. So something is a rented house if the owner is not a natural person and the owner does not live on the house address

```
IF ((Q8<>Q9) AND (Q10<>"N")) THEN sector = H.
```

Any records that can't be filled with the rules we have so far will get the classification **special**. This we have seen in our example for a row of rented houses, the problem with these buildings is that we don't have the inner wall to separate the building into single houses. So to get a full visual representation of all houses will not be possible, but we can classify all addresses present in ACN.

Figure 4: Map of a preliminary result containing all different types of houses.

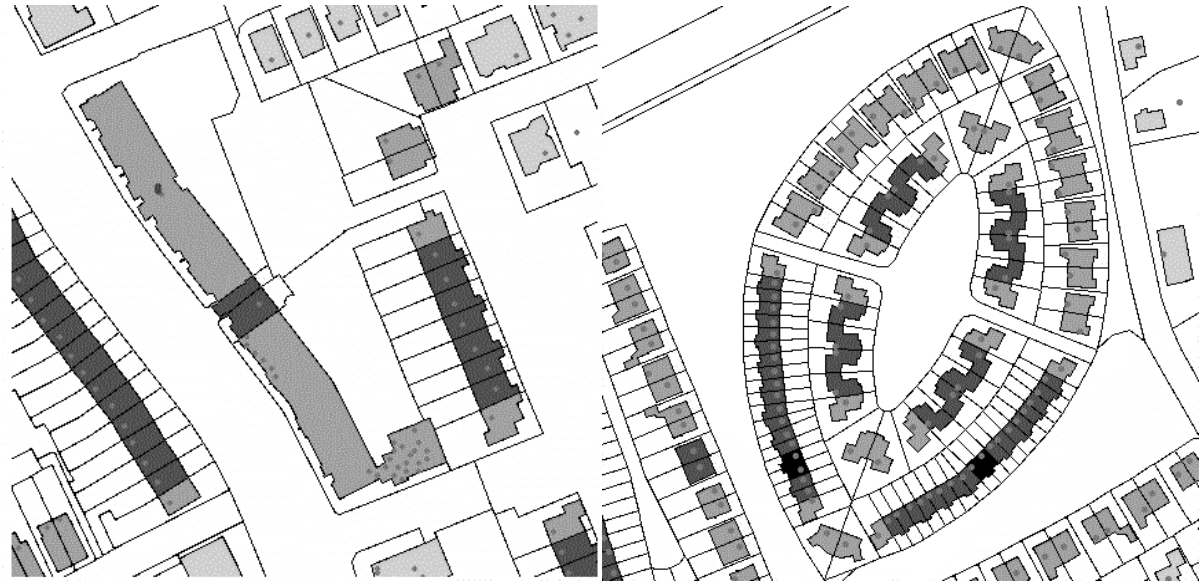

Figure 5: Two more examples from the same map as figure 4.

5. DISCUSSION CONCERNING THE QUALITY OF THE RESULTS

Because the process will be completely automated, an indication of the quality of the results is needed. This will be done in two ways. First, the results of the process will be inspected visually. If there appear to be systematic errors, the classification algorithm should be adapted. During the visual inspection of a preliminary result we found some problem areas that need special attention. In figure 4 we see two rows of houses on one parcel (a little below the centre) which are probable owned by a building corporation. In this case luckily the address co-ordinates (ACN) lay in the houses and can probably help in identification, but the question still remains how we find the inner walls from these kind of buildings. One approach could be using the line in the middle between two address co-ordinates as an internal wall. This approach has not yet been tested.

The second, the classification algorithm will also produce a (per house) confidence value indicating the likelihood of the correctness of the classifications. In normal cases these values will be very high, but the more post processing that is necessary decreases these values. Every type of houses has its characteristics. In our rules we use these characteristics. The amount of characteristics that a house meets to all characteristics could be used as a likelihood for the correct classification.

The classification rules, based on the information on the ownership and the postal address of the parcel, can give wrong results in case this information is not correctly stored in the database or is insufficient. In such cases it is possible that these houses will remain unclassified or get a wrong classification. We can detect these cases for houses that should be classified as one of the middle or end house in a row if we consider the geometrical measurements in our classification. These kinds of houses, in conjunction to the two-under-a-roof type, are normally of the same building-type, so the area and perimeter is more or less the same as the houses in their close neighbourhood. Each unclassified house will be compared with a set of surrounded roughly equally sized and shaped houses. If this set has the same

classification type than the unclassified house will assigned to this value. The confidence value indicating the likelihood of the correctness of the classification can also be based on this rule. In this step we compare each classified house in the same way and the confidence value will be set on the determined correspondence.

Finding the right classification instead of "special"

Any units that can't be classified by the rules defined so far have been given the classification "special". Some of these houses could get another classification in a post-processing step. This post-processing will be discussed now.
Especially the example of "G2" (in figure 3), which is a very common type of houses in the Netherlands, we would like to improve. At this moment this type of house is one of the biggest sources of houses with the classification "special". Characteristic in this case is that there are several ACN co-ordinates inside the building, most of the time neatly distributed. If the number of the address inside the building is more then two and not too big we could use the next steps to separate the building into houses:

- find the centre-line of the polygon using triangulation
- expand the centre-line to the border
- separate equally the expanded centre-line into the number of addresses
- form perpendicular lines on the expanded centre-line at the positions found in the previous step
- expand these lines to the border
- delete the centre-line and rebuild the new formed houses
- If the area of the found houses is within a certain range accept the new formed houses

Give the first and last houses the classification **end-of-a-row,** the others the classification **middle-of-a-row**. This step is only performed to get an as best possible visual representation of the situation.

All other houses that have been given the predicate "special" could get a new classification, using almost the same rules, but losing some of the conditions.
We estimate that over 95% of the houses will get a classification other than special.

6. CONCLUSIONS

In the coming months we plan to prove our ideas described in the previous sections by implementing them. The resulting product will be an important step towards a categorised index of the houses in the Netherlands. After the individual houses (living units) have been computed and classified, a meaningful relationship with the price of a house can be determined. A useful index has to be designed in which the following aspects are taken into account: period (temporal resolution, e.g. month), region (spatial resolution, e.g. zipcode area, municipality). Of course trend analysis and future prediction are the next step.
At the moment of writing this article we have performed the overlay for our test dataset. Our test dataset was given a preliminary classification. We have visually examined this result. We have not yet implemented our solution for rented rows of houses.

ACKNOWLEDGEMENT
We would like to thank the "Kadaster" in the form of the following persons Harry Uitermark, Caroline Groot, Bart Maessen, and Ted Schut for their support, test-data and discussions about this subject.

REFERENCES
[1] ASK-OpenIngres (1994): Ingres/object management extension user's guide, release 6.5. Technical report, 1994.
[2] Fabri, A., G. Giezeman, L. Kettner, S. Schirra and S. Schönherr (1996): The CGAL kernel: A basis for geometric computation. In M. C. Lin and D. Manocha, editors, Proc. 1st ACM Workshop on Appl. Comput. Geom., volume 1148 of Lecture Notes Comput. Sci., pages 191-202. Springer-Verlag, 1996.
[3] Handboek Massale Output (!998): Kadaster Apeldoorn, the Netherlands.
[4] Product Information ACN (1998): Kadata, Dutch Cadastre, February, 1998.
[5] ArcInfo, URL: http://www.esri.com/software/arcinfo

CO-ORDINATES
Maureen Rengelink
Prof.dr.ir. Peter van Oosterom
Drs. Wilko Quak
Ir. Edward Verbree
Delft University of Technology
Faculty of Civil Engineering and Geosciences
Department of Geodesy
Section GIS technology
Thijsseweg 11
2629 JA DELFT
The Netherlands
Tel. +31-15-278 3756
Fax +31-15-278 2745
E-mail oosterom@geo.tudelft.nl
c.q.quak@fgeo.tudelft.nl
e.verbree@geo.tudelft.nl
URL //www.geo.tudelft.nl/GISt/gist_e/index.htm