Clodoveu A. Davis Jr.
PRODABEL -- Processamento de Dados do Município de Belo Horizonte
Av. Presidente Carlos Luz, 1275
30210-000 -- Belo Horizonte -- MG
BRAZIL

ADDRESS BASE CREATION USING RASTER/VECTOR INTEGRATION

Abstract: The city of Belo Horizonte, Brazil, which has a population of slightly over 2 million, has formed a vectorial database for GIS using digital restitution of air photos. Even though this database is very detailed and precise, it contains only the representation of physical objects. The next step taken was the creation of an address database, by placing symbols over the vectorial information, and associating these symbols with the building address. This article describes the processes used to create and validate such a database, by using raster images of older and rather imprecise cadastral maps, which contain all the required information. In the approach chosen, the operator is prompted to locate a given address, and is able to do so very quickly by simply looking at the raster image of the cadastral map, which is displayed under the vectorial information. No typing is required, because the addresses to be located come from an existing alphanumeric database. Notice that high precision on the older map is not required, because it is used only as a reference for the identification of the addresses.

Results obtained include the placement and verification of about 950,000 symbols (400,000 addresses, 270,000 parcel codes and 280,000 tax codes), along with their associated alphanumeric information, consuming 4,700 man-hours, over a period of four months. So, the overall productivity was around 3.4 symbols placed per minute.

INTRODUCTION

Belo Horizonte, the fourth largest Brazilian city, has a population of more than two million inhabitants, spread through 335 square kilometers, and is the center of a metropolitan area that shelters around 3.5 million people. Through a pioneering work in Brazil, Belo Horizonte started in 1989 an effort towards greater economic and administrative efficiency. The objective was to integrate the most significant databases about the city and, through the use of GIS tools, reach the citizens with higher-quality public services and democratic access to information.

The creation of a digital base map was, of course, the first step in this direction. In the beginning of the project, it was decided that it was the time for a major cartographic

update, since the most recent map sheets corresponded to an air photo survey completed in 1972. Even though this material had been suffering minor updates ever since, it was clear that these maps lacked precision and did not reflect the reality accurately.

The project then proceeded, to produce a revised set of landmarks, a new photogrammetry, and a new set of maps. Initially, the creation of maps from air photos was specified in digital form, so that it would always be possible to generate new originals from the files, making the updating process easier. However, the richness of the digital material gathered in this process led to geoprocessing, and a large number of applications were then envisioned to make effective use of the newly collected data.

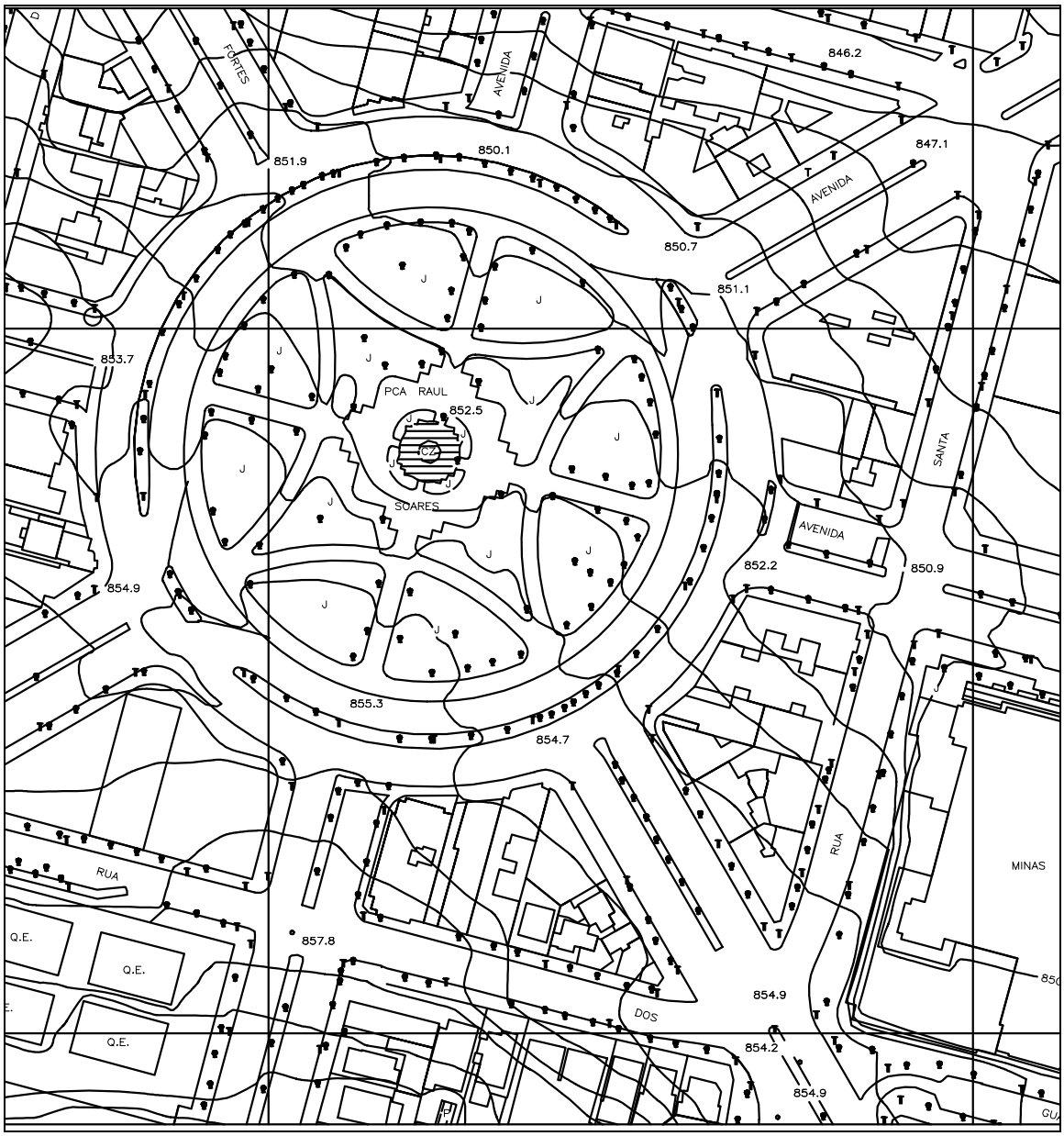

Figure 1 - A sample of Belo Horizonte's base map

The digital database that was created from the air photos includes around 4 million geographic objects, divided in about 90 different categories, or layers. Even though this database is very detailed and precise (figure 1), it lacks the ability of helping users and applications to locate points in the city in a simple and fast way. The next phase would then be the creation of an address base, a necessary and very important step towards bringing other databases to the GIS. The importance of having a good address base was very clear to us, since most existing databases use some kind of addressing. Furthermore, street addresses are one of the main ways that common people have to position themselves in the city.

THE ADDRESS BASE

In Belo Horizonte, given the reality and the great perspectives of the GIS project, a large number of state and federal organizations joined the municipal administration in a cooperative agreement. This agreement meant to reduce data collection and maintenance costs throughout the involved agencies, creating a common set of data for the city, which would be shared by all (Parent 1991).

In the early stages of this agreement, no agency had any hardware or software dedicated to GIS, so efforts were directed towards available alphanumeric data. The first real product of the cooperation was the creation of a common street code table, by means of which addresses from all databases could be compared. Using this big conversion table (Belo Horizonte has about 16,000 streets), it was possible to create an alphanumeric address table, with data collected mainly from the taxing database (maintained by the city's administration), the electric billing database (from the state's power company) and the water billing database (from the state's sanitation company). The resulting alphanumeric address base contains around 410,000 different addresses. After this was done, we gave the address base a final form by adding the corresponding parcel codes and tax codes, according to the city's files. Of course, addresses that came from the other databases would not have any associated parcel or tax code.

The problem now was, how could we georeference the addresses and parcel codes in a practical and economical way? The answer came from observing the existing cadastral maps. Even though they are cartographically imprecise, they contain all of the parcel codes, according to the urban cadastre, and most of the addresses, collected on the field several years ago.

That led us to the process of scanning these older maps, to show them as digital images on the screen of the GIS workstations, underneath the vectorial database (Graça 1990). Using this technique, operators would read the parcel codes and addresses from the raster, in their approximate geographic positions, being then able to decide where to place them in the vectorial base.

Notice that our parcels are not topologically closed yet, because the base map only contained physical objects (walls, fences, buildings), as observed from the air photos. Not

having closed parcels shuts the door to many interesting applications, but we figured that there were other ways to get the GIS data to good use quickly, without so much effort. So, we decided to represent addresses, parcel codes and tax codes as different symbols, that are related to one another through the GIS data model, saving the parcel closing problem to a later stage. When we finally get our parcels closed, the correct codes will already be placed, and the symbols we created could then be treated as centroids for the parcel polygons.

Not much precision was required, because the symbols could be placed anywhere inside the parcel, for future uses. In this sense, the older maps' lack of precision did not matter much, but their use in raster format would lead to significant productivity gains.

RASTER BASE FORMATION

Since the GIS software we use only adds the capability of *displaying* raster images underneath vector data, but does not include any tools to *prepare* raster images for display, we had to develop some algorithms of our own.

Our GIS software demands that only the inner part of the map (i.e., the map itself, excluding margins and legends) must be tied to a vectorial rectangular object, which has to contain an alphanumeric field to indicate the raster file name. This rectangle should be placed in the geographic coordinates that correspond to the part of the space covered by the map sheet. When it is necessary to display the raster file, the software looks for the correct filename within the rectangle's attributes and then displays it inside the vectorial bounds. Using this approach with articulated map sheets, that cover the whole city, we have a *continuous raster database*, also interesting for many other applications.

The tricky part is how to remove the unnecessary information from a scanned map sheet, and, preferentially *at the same time*, execute the proper adjustments to the image, in order to compensate for scanner skewing and other distortions. Our algorithm uses a particularization of the resampling technique (Niblack 1986; Gonzales and Wintz 1987; Davis 1992), demanding from the user the coordinates of the four corners of the inner part of the map, to be used as control points. Using the control point coordinates and the size of the resulting image, it is possible to, simultaneously, correct the image geometry and save only the interesting part of the map. The coordinates of the corners, i.e., the control points, are obtained interactively, using the mouse, through a specially written program.

Figure 2 shows a diagram of the technique, which corresponds to an *output-to-input resampling*. Initially, a transformation function is determined, using the correspondence between the control point coordinates (A, B, E, F), from the original image, and the image extremities (H, J, K, M), from the transformed image. For each of the pixels (like D, in the diagram) in the output image (whose size -- a by b pixels, in the diagram -- is defined by the user), a pair of coordinates (xP, yP) in the input image is then calculated, using the transformation function. Most commonly, these coordinates will not be round numbers,

making it necessary to interpolate, in order to determine whether the pixel in the output image should be black or white. Of course, because the images are always binary, the interpolation can be easily substituted by simple IFs. The algorithm's performance is further enhanced by using proper buffering and memory-handling techniques.

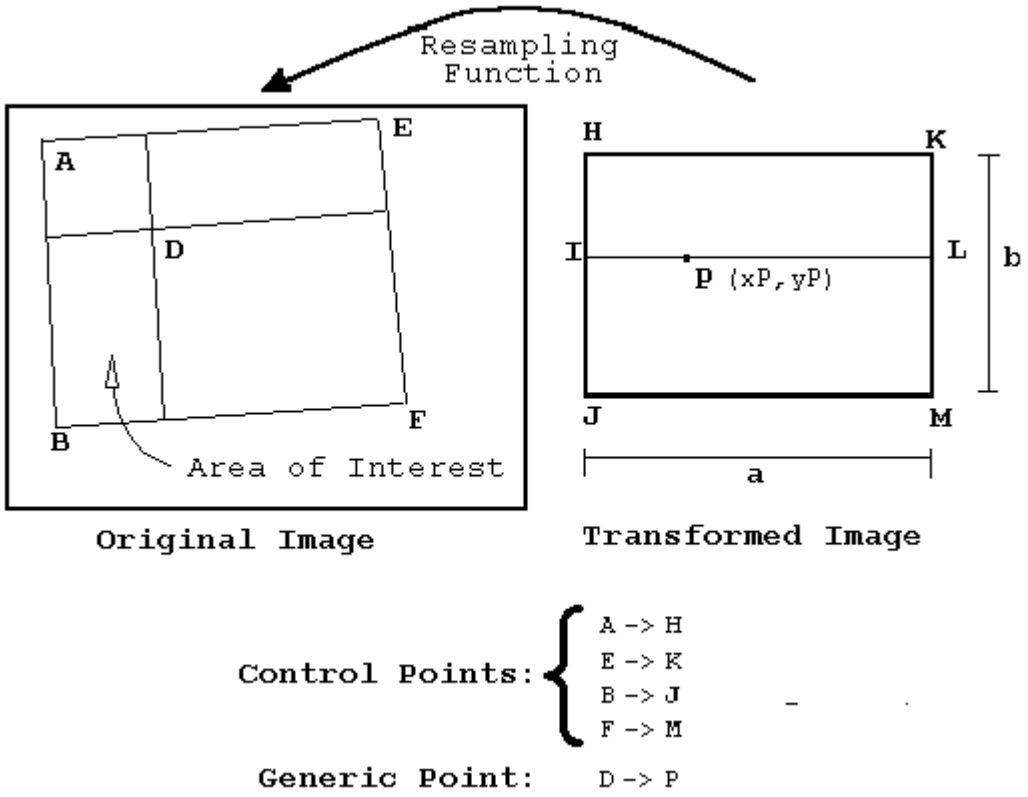

Figure 2 - Resampling diagram

For our application, we had to use this algorithm with 1,226 A1 map sheets, our 1:1000 scale set. As a reference for future study of the effects of scanner distortion and skewing, we include here Table 1, which shows the variations in length and rotation of the four sides of the rectangle that contains the area of interest, in the raw scanned maps. The observed side lengths and rotations have been calculated from the control point coordinates. Since the scanning resolution is 200 dots per inch, we can calculate the expected lengths in pixels for the "ideal" rectangle. All sheets cover 600 by 500 meters of terrain, which corresponds on paper to 600 by 500 millimeters.

As it can be seen, the linear distortions are in the range of 7 ± 10 pixels, which would correspond, in this case, to an extreme distortion of about 2.2 meters, or 2.2 millimeters on paper, which is much less than the actual imprecision contained in these maps. The rotation, which was thought of as a big potential problem, in fact is not so significant.

TABLE 1
OBSERVED SCANNER SKEWING AND DISTORTION

	Length (pixels)			Rotation (degrees)	
Side	**Expected**	**Average**	**Std.Dev.**	**Average**	**Std.Dev.**
Left	3937	3952	2.9	0.20	0.16
Right	3937	3946	10.7	0.25	0.31
Top	4724	4729	13.1	0.21	0.19
Bottom	4724	4727	12.7	0.19	0.16

ADDRESS PLACEMENT

With the rasters at hand, and having the alphanumeric address base, we proceeded to create the actual address placement routines. There was a reasonable fear of imprecision in the address base, too, since the expected figure for the city's addresses was something around 310,000. So, there was a probable 100,000 erroneous addresses in the alphanumeric database, a number that was allowed to grow that high mainly because of the impossibility to consist address data. If postmen could live with that (while delivering taxes, electrical and water bills), so had we. The placement routines would have to be made smart enough to take care of special cases that might occur, without breaking the speed of the operation.

To do that, we added a field to reflect the quality of the address. This field would contain:

- **OK**, if the address can be confirmed visually using the raster information;
- **Probable**, if the address cannot be found on the raster map, but can be easily identified as a simple mistake: for instance, swapping two numbers in the street number;
- **Not found**, if the address could not be placed by some reason. In this case, the operator is prompted for a description of the reason why the address could not be found. Common explanations include "out-of-range numbering", "irregular numbering", and "street not found." At a later stage, many of those addresses will have to be field checked.

To further enhance the speed of the address placement, we separated the alphanumeric address base by blocks, according to an existing classification, used by the parcel codes. We then generated a large number of small ASCII files, containing addresses by block, sorted by street code and street number. In these files, each address can be related to a parcel code and a tax code, according to the existing alphanumeric

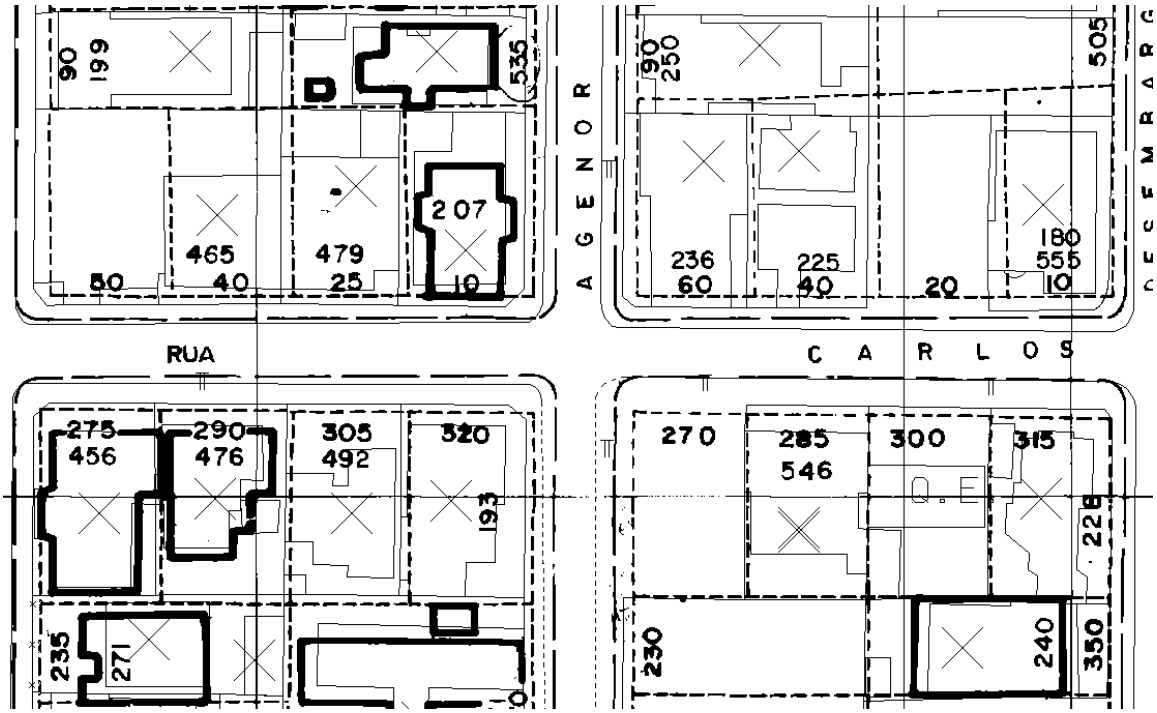

Figure 3 - Example of raster/vector integration
(The X's correspond to address symbols)

data. In most cases, we were placing not only a point, but three, georeferencing tax and parcel codes along with the addresses. Symbols that got placed simultaneously were also tied by means of internal database relations, so that we could obtain the parcel code from a given address, and other types of combinations. This led us to the final form of the placement routine, which would be:

1. Select a block, indicating its code (annotated on the raster);

2. Start the placement routine, which will ask for the block code;

3. Place a block symbol somewhere inside the block (as with parcels, our blocks are not yet topologically closed);

4. The routine will open the external ASCII file containing addresses for that block, as gathered from the alphanumeric databases;

5. In order, each of the addresses will be presented to the operator, requesting the location, to be confirmed using the raster;

6. If the address can be placed, click the left mouse button in the correct geographic position. The routine then prompts the operator to choose a quality status from a menu (**`OK`** or **`Probable`**, as explained later);

7. If the address can not be placed, click the right mouse button. The routine will prompt the operator for a textual reason for not placing the address;

8. Loop to step 5, until the end of the block.

The routine includes mechanisms for interrupting the work in the middle of a block, and to resume it later without duplicating work. Some other consistency routines were developed later, so that operator mistakes could be catched.

Figure 3 shows an example of raster/vector overlay. Notice the parcel codes and street addresses annotated on the raster, and the differences between the thin vectorial lines and the thicker raster lines. Figure 4 shows a sample of the final results, with street numbers in the place of the address symbols.

figure 4 - Final results sample

RESULTS

Table 2 lists some statistics regarding the operation of the described technique. It has to be taken into account that our operators were not acquainted at all with computer equipment, being urban cadastre technicians and draftsmen. Their training, which had to include basic keyboard and mouse operation, was conducted in three days, followed by a two-day hands-on training. From then on, they were in actual production. None of these operators had ever performed a similar job, neither using GIS nor any other graphic system. What counted most, in this case, was their experience with the city's traditional cartography: many of them can easily tell which region is on the screen simply by looking at the shapes of the streets.

TABLE 2
ADDRESS PLACEMENT PRODUCTIVITY

Total number of productive days	109
Number of operators	9 during 59 days; 10 during 50 days
Total operational time	6,483 h
System stops	1,697 h (26,2 %)
Productive time	4,786 h (73,8 %)
Addresses placed	401,000
Parcel codes placed	269,133
Tax codes placed	279,052
Blocks placed	14,284
Average addresses per minute	1.4
Average symbols per minute	3.4

CONCLUSIONS

The overall productivity number, of 3.4 points per minute, reflects the method's efficiency. It is our opinion that further enhancements could be achieved if the operators had had previous experience with the GIS system, as they do now. But it is unquestionable, to us, that this productivity could not have been achieved if we did not use the raster/vector integration technique. At a very low cost, the use of raster/vector integration led us to gain also in terms of consistency, since the risk of operator

placement mistakes was greatly reduced. We are currently using a similar integration to create the city's street centerline network.

Furthermore, looking back on the experience, we found out that perhaps georeferencing simultaneously three different informations (addresses, parcel codes and tax codes) has complicated the operation a bit. The placement of a single information could lead to better productivity results.

Time wasted on system stops reflects our own lack of experience with the GIS software at the beginning of the project. Today, this number has retreated to an acceptable minimum.

REFERENCES

Davis Jr., Clodoveu A., 1992, *PixelWare: A Digital Image Processing System* (in Portuguese), MSc Thesis, Belo Horizonte, MG, Brazil: Federal University of Minas Gerais

Gonzales, Rafael C.; Wintz, Paul, 1987 *Digital Image Processing*, 2nd. Ed., Reading, MA: Addison-Wesley Publishing Company

Graça, Lúcio M. A., 1990 "Using Scanners for Digitizing Topographical Charts and for Implanting an Urban Geographic Information System" (in Portuguese) in *Proceedings of the First Brazilian Symposium on Geoprocessing*, São Paulo, SP, Brazil: University of São Paulo pp. 219-224

Niblack, Wayne, 1986 *An Introduction to Digital Image Processing*, Englewood Cliffs, NJ: Prentice/Hall International

Parent, Phil, 1991 "Cooperation: a Key to System Success" in *GIS World*, vol. 4, number 8, November 1991, pp. 28-29

ACKNOWLEDGMENTS

The author wishes to thank the Foundation for Research Support of Minas Gerais (FAPEMIG), for granting the necessary funds for the presentation of this paper.

The author also wishes to thank PRODABEL's GIS professionals, declaring that this paper is a truly an achievement of the whole team.