IBGE Statistical Grid in Compact Representation

Peter Krauss[1], Luis Felipe Bortolatto da Cunha[2], Thierry Jean[1]

[1] Instituto de Tecnologias Geo-Sociais AddressForAll
Av. Paulista, 171 – 4° andar – Bela Vista – São Paulo – SP – Brasil

[2] Universidade Federal do ABC (UFABC)
Alameda da Universidade, s/n° – Anchieta – São Bernardo do Campo – SP – Brasil

{peter,thierry}@addressforall.org, luis.cunha@ufabc.edu.br

Abstract. *This article describes the development of the IBGE Statistical Grid in Compact Representation, an alternative structure to the original grid that aims to improve its use in databases and enable new applications. It was implemented in the PostgreSQL+PostGIS environment, and its main advantages are direct indexing by cell geocode and reduction of disk occupancy, both during operation and during package distribution. A library of functions was made available along with the distribution, solving the encoding/decoding of cell geocodes through simple "snap to grid". Future work includes developing a similar grid with Geohash-like indexing, making geocodes shorter and hierarchical.*

1. Introduction

Grid Systems are a regular-sized geospatial data structure that allows detailed analyses independent of political-administrative or operational territorial divisions, while also meeting the need of storing data in small and stable geographic units over time and facilitating data from multiple origins and types (e.g., vector and raster) to be integrated in the same format. According to Bueno (2016), standard grid systems advantages include: (1) spatiotemporal stability; (2) adaptation to spatial cutouts; (3) hierarchy and flexibility; (4) versatility; (5) cartographic interpretation; (6) simple identification; (7) use in modelling; and (8) minimization of MAUP[1] effects.

In 2016, the Brazilian Institute of Geography and Statistics (IBGE) made available a grid system covering the entire national territory, named Statistical Grid, composed of 7 hierarchically coupled grids and population information, which is the official grid system of Brazil. This product became possible due to technological advances adopted in the years prior to the 2010 Population Census, such as the use of electronic collection devices that could capture geographic coordinates and the development of an address database connected to the road mapping (IBGE, 2016). The IBGE Statistical Grid included selected data from the previous census which provided a significant increase in detail, particularly in rural regions, compared to previous data dissemination methodologies.

The IBGE Statistical Grid introduced significant advances, including a Coordinate Reference System (CRS) that can cover the entire Brazilian territory in a constant-area regular-sized grid and a methodology for aggregation and dissemination

[1] The Modifiable Areal Unit Problem (MAUP) is a source of statistical bias that arises from the choice of geospatial data aggregation unit.

of population census data in a grid system. But its distribution format and lack of technical documentation have limited its applications.

This paper describes an ongoing research that aims at improving access to the IBGE Statistical Grid and making it more computationally efficient. It proposes an alternative structure, named Compact Representation, that can be implemented in SQL database and have the main advantages of:

(1) Indexing the grid directly by its geocodes: the internal database cell identifier, *gid*, can be the cell's name (a numeric geocode), making search and retrieval operations much simpler and faster.
(2) Improved search and retrieval operations, due to the geocode indexing.
(3) Reduced distribution size, from 849 Mb (56 zip files) to one zip file of 47 Mb.
(4) Reduced SQL database disk usage to approximately 20% of the original size.
(5) Distribution in a non-proprietary open format, the CSV, a universal open standard that can be opened in any data management and analysis software, unlike the original Shapefile format.
(6) A utility kit of optimized algorithms, including encode/decoding, snap-to-grid and drawing cells.
(7) Easy modularization, providing data fragmentation and main functional modules to any simple SQL database (e.g., SQLite in Android-OS), with no need for GIS extensions.
(8) Caching the aggregate non-geometric data of all parent-grids.

This research also reveals some aspects about the grid that had to be obtained by reengineering, given the lack of technical documentation, such as the relationship amongst each cell and its cell name (ID). Although it is a faithful and complete reproduction of the original grid, some decisions are arbitrary (e.g., the use of *gid* instead of the original *id*). All development was done in PostgreSQL+PostGIS environment, and the application source code is available as Git repository at http://git.osm.codes/BR_IBGE.

The changes made to the statistical grid should also enable its use as a multipurpose geocode system (geohash). A geocode can express approximate geographic coordinates in a unique identifier, which is usually small and human readable. Geocodes can be used for labeling, data integrity, geotagging, and spatial indexing (KRAUSS et al., 2020; KRAUSS & ALMEIDA, 2020).

The remainder of the paper is structured as follows: Section 2 and Section 3 describes the Original Grid and Compact Representation specifications, respectively, Section 4 includes a short description of the developed algorithms and Online API, and Section 5 concludes the paper, also indicating future research pathways.

2. Original Grid specifications

The IBGE Statistical Grid is a hierarchical grid system built over Albers Equal-Area Conic Projection and SIRGAS2000 horizontal datum, with the main characteristic of area equivalence. The Coordinate Reference System (CRS) main parameters are specified in Table 1 (IBGE, 2016).

Table 1. IBGE Statistical Grid CRS-Albers specifications

Parameter	Specification
Standard parallel 1	-2
Standard parallel 2	-22
Latitude of origin	-12
Central meridian	-54
False Easting	5000000
False Northing	10000000
Unit	Meter

Brazil territory was fully covered by 56 squares of 500 km side, and subsequently divided six times into squares with sides measuring 1/5 or ½ its previous size to form the next grid, with lower scale and higher resolution, as shown in Figure 1.

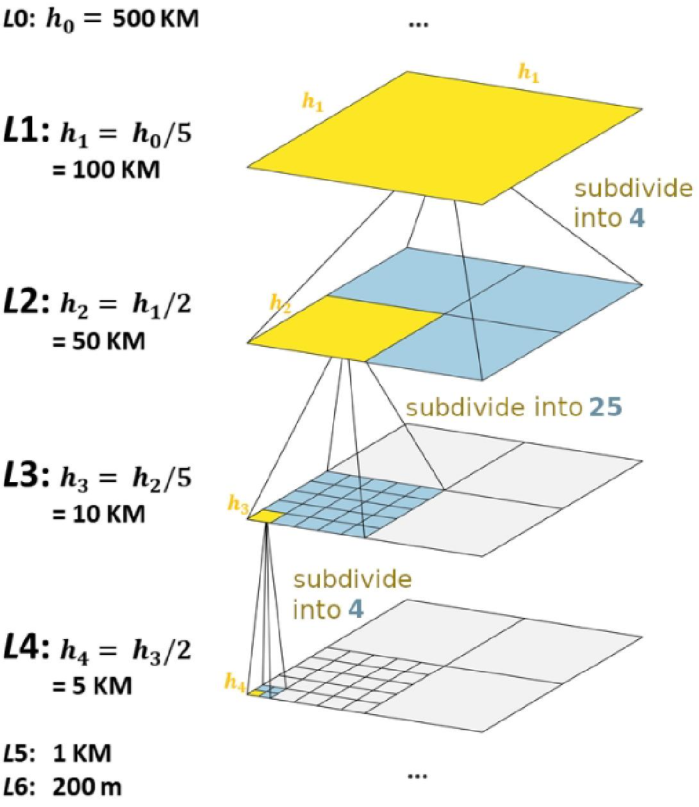

Figure 1. IBGE Statistical Grid levels

On *L*5 and *L*6 geometries (1 km and 200 m, respectively) relevant data from the 2010 Population Census were added, with L5 being the default level for rural areas and L6 for urban areas.

The original ID of any cell is generated following the template: "*{side}E{X}N{Y}*", where *{side}* equals the cell side size (200m, 1 km, 5 km, 10 km, 50 km, 100 km or 500 km), and *{X}* and *{Y}* equals the coordinates of the cell having its corner as reference — upper right, excluding the last 2 digits, for 200m cells and lower right, excluding the last 3 digits, for all other levels.

The "ID_UNICO" column refers to the geometry column and is the smallest cell of an area (200 m for urban areas and 1 km for rural areas). All other "nome_" columns ("nome_1km", "nome_5km", etc.) are parent-cell references.

Table 2. IBGE Statistical Grid structure

Variable	SQL Type	Comments
ID_UNICO	varchar(50)	Cell's unique identifier
nome_1km	varchar(16)	(redundant) L5 id
nome_5km	varchar(16)	(redundant) L4 id
nome_10km	varchar(16)	(redundant) L3 id
nome_50km	varchar(16)	(redundant) L2 id
nome_100km	varchar(16)	(redundant) L1 id
nome_500km	varchar(16)	(redundant) L0 id
QUADRANTE	varchar(50)	(redundant) Alternative L0 id
MASC	integer	Male population
FEM	integer	Female population
POP	integer	Total population
DOM_OCU	integer	Occupied households
Shape_Leng	numeric	(redundant) Shape length
Shape_Area	numeric	(redundant) Shape area
geom	geometry	Cell geometry

3. Compact Representation specifications

Most of the variables made available in the original database are redundant and could be summarized in a more compact form. Tables 2 and 3 show the structure of the original and compact representation.

Table 3. IBGE Statistical Grid in Compact Representation structure

Column	SQL Type	Comments
gid	bigint NOT NULL PRIMARY KEY	New unique cell identifier
pop	integer NOT NULL	Total population
pop_fem_perc	smallint NOT NULL	Female population percentage
dom_ocu	smallint NOT NULL	Occupied households

On the Compact Representation, the unique ID ("ID_UNICO") was simplified and referred to as *gid*, which structure is as follows for human-readable decimal: "*{X}{Y}{L}*", where *{X}* and *{Y}* refer to the complete coordinates of the reference-point in the corner of the cell, always containing 7 and 8 digits respectively, and *{L}* refers to the level (1 digit).

The *gid* column compresses all the cell reference point location information into a single 64-bit integer and allows the indexation, not only of the smallest cell of an area, as the Original Representation, but of all the others, creating a large and economical cache of summarization grids.

The cell geometry is not stored in the Compact Representation, but it can be quickly reconstructed in PostGIS from the *gid* and CRS-Albers specifications. The changes made to the Statistical Grid reduced the distribution size from 849 Mb (56 zip files) to 1 zip file of 47 Mb. It also reduced the disk usage in the SQL database in 83%, and the *gid* as index made the proposed algorithms more responsive.

4. Algorithms and Online API

The research repository (Git) includes source codes for installing the Compact Representation distribution in PostgreSQL+PostGIS environment, setting up an Online API, conversion between the Original Grid and Compact Representation, and optimized algorithms for manipulating these data. The algorithms include:

- **Encoding/decoding**: the conversion between id, gid and geometry demands a certain level of synthetic control. There is a set of functions that deals with these conversions.
- **Snap to grid**: instead of using the PostGIS geometric operations, discretization functions are responsible for identifying in which cell a point is contained. This process includes the conversion from point coordinates (WGS 84) to the CRS-Albers of the grid, and the identification of the cell.
- **Drawing cells**: allows the visualization of the grid, using *gid* drawing functions.

Some library functions work in different coordinate systems, but all are internally standardized, following the conventions:

- **LatLon GeoURI**, (*lat,lon,uncert*) where *lat* is latitude, *lon* is longitude and *uncert* is uncertainty (radius of the uncertainty disk in meters).
- **Albers XY**, (*x,y,Level*) where *x* and *y* are the Albers coordinates and *Level* the hierarchical level of the grid.
- **Unit IJ**, (*i,j,s*) where *i* and *j* are indices of the "unit square grid" and *s* is the size of the cell side of this grid.

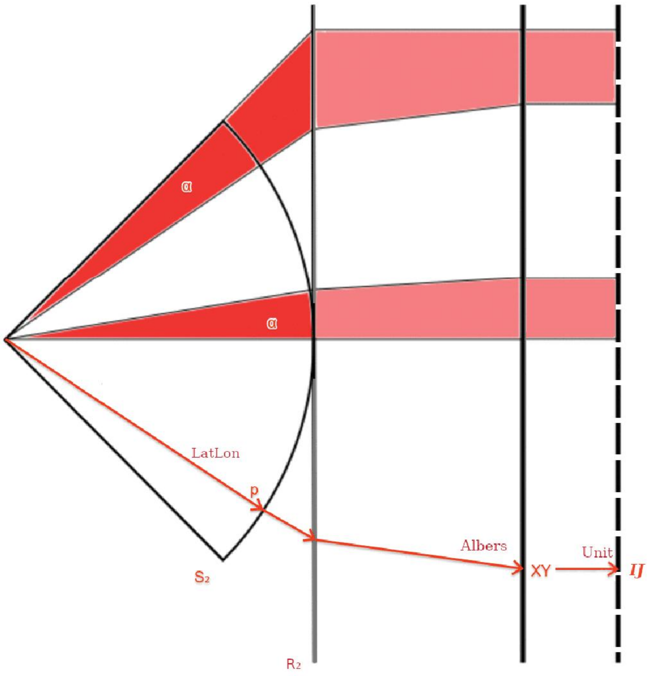

Figure 2. CRS conversion description

The Online API is based in the Geo URI internet standard (RFC 5870 of June 2010), that offers a simple and consistent interface to return information from a point or its neighborhood. When the user input any LatLong point at Brazilian territory, the API returns the attributes of the 1 km cell that covers that point. For instance, the location with coordinates 15°48'S 47°51'W and Geo URI standard "geo:-15.8,-47.86;u=500" (where u equals uncertainty in meters), can be accessed at our endpoint, http://osm.codes/geo:-15.8,-47.86. Accessing the information of a location by the cell geocode would be possible using a Geo URI expansion that establishes consistent conventions (Krauss et al., 2020). Using the proposed syntax, a request by cell name would be “geo:BR_IBGE_2010: 1KME5649N9566”, although it is a feature not yet implemented.

5. Conclusions and future work

This ongoing research is mainly concerned at improving access to the IBGE Statistical Grid. It proposed an alternative compact structure that reduces the distribution size and database disk usage, while functionally reproducing the original grid data structure. The PostgreSQL+PostGIS application also includes optimized algorithms for encoding/decoding, snap to grid, drawing cells and an Online API. Portability to ArcGIS framework is also planned.

Future work includes the development of a cell coverage algorithm, that could rapidly return the attributes of an area, such as absolute or relative population estimate. A detailed assessment of the Compact Representation as a multipurpose geocode system should also be carried out, although an early evaluation suggests that further adaptations to its structure and naming scheme might be necessary (KRAUSS, 2021).

References

Bueno, M. D. C. D. (2014). “Grade estatística: uma abordagem para ampliar o potencial analítico de dados censitários” (doctoral thesis, UNICAMP, Campinas, Brazil).

IBGE – Instituto Brasileiro de Geografia e Estatística (2016). “Grade Estatística”. Retrieved from https://geoftp.ibge.gov.br/recortes_para_fins_estatisticos/grade_estatistica/censo_2010/grade_estatistica.pdf.

Krauss, P., & Almeida, R. (2020). “Grade estatística do Brasil: uma proposta de melhora orientada a geocódigos hierárquicos e multifinalitários”. “II Simpósio Brasileiro de Infraestrutura de Dados Espaciais”, Brazil. Retrieved from https://inde.gov.br/images/inde/poster1/NovaGradeIBGE-poster-v2.pdf.

Krauss, P., Jean, T., & Bortolini, E. (2020). “Expansão do protocolo GeoURI (RFC 5870 da internet) visando a interoperabilidade de geocódigos nacionais soberanos”. “II Simpósio Brasileiro de Infraestrutura de Dados Espaciais”, Brazil. Retrieved from https://inde.gov.br/images/inde/poster3/Expans%C3%A3o%20do%20protocolo%20GeoURI.pdf.

Krauss, P. et. al. (2021). “Grade Estatística IBGE em Representação Compacta”, Git repository at http://git.osm.codes/BR_IBGE.

Krauss, P. (2021). “Geohash adaptado à Grade Estatística IBGE”, Git repository at http://git.osm.codes/BR_IBGE_new.