SPATIAL DISTRIBUTION QUERY LANGUAGE

Piotr Bajerski[1] and Jacek Frączek[2] and Dariusz Mrozek[3]

Department of Computer Science, Silesian University of Technology

ul. Akademicka 16, 44-100 Gliwice, Poland

Email: [1]bajerski@zti.iinf.polsl.gliwice.pl, [2]jacekf@polsl.gliwice.pl, [3]mrozek@zti.iinf.polsl.gliwice.pl

ABSTRACT

Spatial analysis in Earth Sciences often concerns continuous distributions in space. In the paper the term spatial distribution denotes a set of geographical observations representing values, behaviour or characteristic of a particular phenomenon across many locations on the surface of the earth (e.g., distribution of airborne suspended matter over Poland). Spatial distribution is used as the synonym of a field (geofield). In the case of point measurements, lots of measurement data sets are collected, and then usually stored in a relational database. Traditional systems use script languages and different versions of Tomlin algebra for processing continuous fields using raster representation. The main idea of the work is to extend a query language in such a way that all conditions on spatial and non-spatial attributes of continuous phenomena and discrete objects may be written in a single query. Such a declarative notation makes it possible to globally optimise the query on the server. Sophisticated optimisation of queries is possible when the following assumptions are set up: (1) values of the phenomena distribution for any location can be evaluated on the collected measured values by means of interpolation, (2) queries concern disjoint intervals of distribution values, (3) computations are carried out in discrete space with a resolution specified by a user.

The paper presents requirements, assumptions and the syntax of a Spatial Distribution Query Language (SDQL). The SDQL is an extension of standard SQL that allows to formulate queries concerning spatial distributions. The SDQL is used in an experimental system called Spatial Distribution Server (SDS) (Bajerski, 2000). SDS processes spatial data using linear quadtree ordered by N-Peano fractal space filling curve.

1 INTRODUCTION

Geographical space perceived by human beings may be conceptually modelled as a composition of distributions of continuous phenomena (fields) and discrete objects (Laurini, 2004). There has been a lot of research devoted to modelling and processing discrete entities. The widely used point/line/area approach has become the foundation of many commercial implementations of GIS (Rigaux, 2002). Traditional approach for field processing is based on raster representation of data and a form of Tomlin algebra called map algebra. This method is very flexible, but represents procedural approach, which does not permit optimisation of the whole analytic task. Declarative description of the task – with the use of just one query involving compound conditions on spatial distributions, other spatial data and on non-spatial data – makes it possible to consider the characteristic of all data types and optimise the query globally. Fields approximations may be dynamically created during query execution or generated in advance and then stored in a database. The process of distributions generation may easily dominate the whole cost of query execution due to a large number of

interpolations. Therefore such queries demand for the application of special optimisation techniques.

Our proposal is to extend the syntax of standard SQL to submit queries concerning continuous geophysical fields. It was assumed that specification of the extensions should follow the *OpenGIS Simple Features Specification For SQL* (OGIS, 1999) and object-oriented extensions defined in SQL99. The presented solution is the superset of the standard SQL based on the *select-from-where* framework. Such an approach facilitates retrieval of spatial and non-spatial data in a uniform well-known standard way.

On the assumptions that: (1) values of the phenomena distribution for any location can be computed using measured values by means of interpolation, (2) queries concern disjoint intervals of distribution values, (3) computations are carried out in discrete space with a resolution specified by a user – the approximation of a spatial distribution may be considered as a specific kind of spatial data that consists of a set of discrete area objects with complex borders and many holes. Using quadtree ordered by a space filling fractal curve makes it possible to approximate a distribution in a top-down fashion (Bajerski, 2002) eliminating a significant part of interpolations in comparison to raster representation and avoid using computational geometry during operations on fields (Laurini, 1994). Fig. 1 illustrates the idea of approximation of a hypothetical distribution represented in the form of measurements in scattered points (a), by means of isolines (b), and a quadtree (c), assuming that the user set up in the query interval boundaries *2*, *5*, *9*.

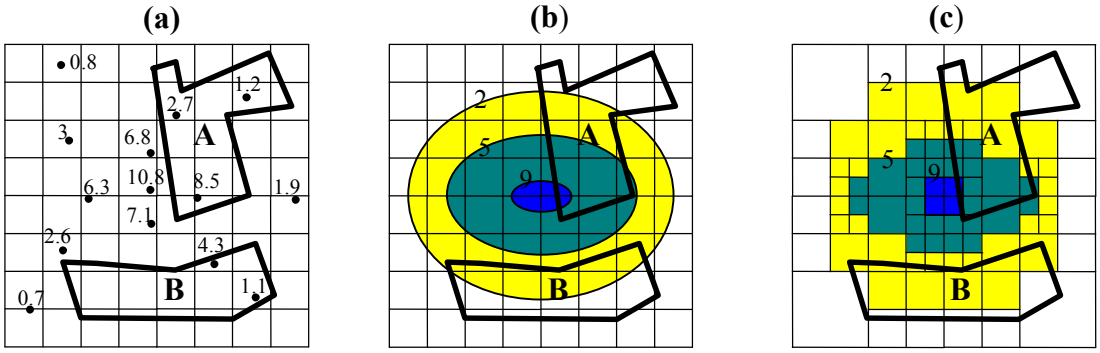

Figure 1. Geofield with discrete objects A and B overlaid: (a) measurement points, (b) isolines approximation, (c) quadtree approximation split to 5th level.

A two-stage algorithm for approximation of a geofield by means of quadtree was proposed (Bajerski, 2002). In the first step squares are split according to the values in the measurement network nodes, so the squares' area contains no nodes or has nodes, which values belong to one interval. In the second step this grain approximation is refined using interpolation in selected points and spatial continuity model of the distribution, used to predict where the geofield value may intersect interval boundary. The proposed model of distribution continuity is based on analysis of distribution of differences between measurement network nodes according to the distance between them. Parameterization of the model allows the user to decide whether he wants to obtain less accurate answer faster or wait longer for more accurate one. The method gives a significant speedup in comparison with computations on raster.

As the description of a geofield approximation method is complex, geofield metadata was introduced in the system. A user can save information how to reconstruct a geofield in metadata and reference to it later in a submitted query. For a specific geofield, the metadata defines: square classification method, interpolation method and corresponding parameters as

well as optional context area for which a given geofield may be recreated and default intervals.

The paper is organized as follows: in the next chapter we discuss related works. Then the geofield query language assumptions and syntax are briefly described. In the fourth section appropriate examples are presented.

2 RELATED WORKS

The basic requirements for the spatial query language are introduced in (Egenhofer, 1994). The representative examples of scientific-based proposals for spatial data processing languages are Phenomena (Paulino, 2003; Laurini, 2004) and GeoSQL (Wang, 2000). Particularly, the Phenomena language concentrates on geofields, however it does not allow to include in a query the information how to create a continuous field based on point values stored in a database. Camara *et al*. in (Camara, 1995; Camara, 2000) discusses problems of formal definition of operations on fields and discrete objects and integration of these two approaches in spatial analysis. However there is no reference to query optimisation issues. Simple Feature Specification for SQL (OGIS, 1999) of the OpenGIS consortium defines a framework for processing spatial data in vector representation.

The leading commercial GIS solution, ESRI ARC/INFO, allows to create and process fields in raster format using map algebra. Oracle's extension for spatial issues, called Oracle Spatial, is oriented towards vector model and remote sensing and does not provide the possibility to reconstruct continuous fields.

3 GEOFIELD QUERY LANGUAGE

The syntax of the geofield query language is based on the standard SQL *select-from-where* framework and generally follows OpenGIS (OGIS, 1999) specification of spatial operations and topological spatial relation predicates. As the size of the paper is limited, only main assumptions of the language and its characteristic features are presented:

- New extended constructs should be a superset of the SQL standard.
- The query syntax does not cover map formatting options – edition is done in later stages.
- A query may concern: (1) intervals of continuous geofields, (2) discrete spatial objects, and (3) non-spatial data.
- Continuous scalar fields created from measurement points are approximated by areas in which field values belong to intervals specified in a query.
- Approximations of geofields can be dynamically created during query execution (appropriate parameters are specified in a query) or can be pre-computed earlier and stored in a database. Geofields are treated as tables of a predefined structure.
- Operations like *Distance*, *Buffer*, *ConvexHull*, *Intersection*, *Union*, *Difference*, and *SymDifference* accept as arguments: geofields zones (areas of the field belonging to selected intervals) and discrete objects geometries.
- The subset of topological spatial relation predicates chosen for processing geofields includes: *Equals*, *Disjoint*, *Intersects*, *Touches*, *Within*, *Contains*, and *Overlaps*.

3.1 Select Clause

The *select* clause defines the form and structure of information that is returned to a user as the result of a query execution. The proposed extensions to the *select* clause in the spatial distribution query language provide the possibility to:

- Return the approximations of geofields – in the form of areas belonging to intervals that comply with the query conditions.
- Return the boundaries of spatial objects in the form of a vector representation.
- Operate on spatial attributes (intersection of geofields, etc.).
- Use aggregate functions with spatial attributes.

The result of a query takes a form of a map or a table, depending on attributes and functions used in the query. The use of spatial attributes leads to a map presentation. In this case any other non-spatial attributes are displayed over the map's base.

3.2 Referencing Geofields

Geofields appearing in queries are treated as tables of a pre-defined structure containing the following pseudo-columns:

- *Zone* – the area belonging to a given interval of a phenomena's value or the area created as the result of operations on regions belonging to intervals fulfilling the query conditions. A *zone* may consist of many disjoint parts.
- *Value* – the value of a geofield in a specific point.
- *Interval* – the identifier of a distribution value range, to which a raster cell was classified.
- *X, Y* – geographical coordinates of a point in the context area.

3.3 Parameters of Geofield Reconstruction Process

To dynamically create the approximation of the indicated geofield the following must be defined: (1) measurement data set, (2) square classification method, (3) the continuity model (required for selected classification methods), (4) interpolation method with suitable parameters (semivariogram for Kriging, the exponent for Sherpad's method, etc.), (5) methods for selection of measuring network nodes used in interpolation process (quarters or eighths, range radius, number of neighbours). In the geofield generation process, it is possible to specify the boundaries of intervals (by the use of the *intervals* phrase), which are later referenced in queries.

3.4 Where Clause

The syntax of the *where* clause was extended to allow imposing conditions concerning geofields:

- Explicit conditions for field values or values interval identifiers, defined in the form of: comparison with a constant value, comparison with values of another distribution, or examination, whether a value is defined (*is [not] null*).
- The area for which geofields are reconstructed and discrete objects are retrieved is called a *context area*. The context area may be defined in the *where* clause implicitly as a union of shapes of discrete objects referenced in a query or explicitly, with the use of the *context_area* pseudo-column (as a rectangle, a polygon, or a subquery).

- The obligatory *resolution* phrase specifies the size of the elementary approximation square (in distance units or by the number of squares the longer side of the context area is divided by).

4 QUERY EXAMPLES

The section contains examples illustrating the usage of the proposed syntax. The examples use the ambient air pollutant data from Upper Silesia gathered by District Sanitary-Epidemiological Station, Katowice, Poland.

If a user wants to find out areas where the average concentration of airborne suspended matter in 1989 belonged to a certain range, and he may submit the following query:

```
Select zone From ASMAvg89DistrECK
Where value > 110 and value <= 165 and resolution = 2000
```

The query implicitly uses the definition of the reconstruction of the average distribution of suspended matter in 1989 (over the default context area) stored in the metadata under the name *ASMAvg89DistrECK*. The query returns a map showing areas on which distribution values fall into the range: *(110-165]*. If the user wants to display information concerning more intervals, he may write many conditions using the *value* pseudo-column or use *Intervals* phrase and the *interval* pseudo-column. The *Intervals* phrase allows to explicitly define required boundaries of intervals. The *interval* pseudo-column allows to reference the earlier defined intervals by the specification of their number identifiers.

If the requested distribution is not available in metadata, a user may create the distribution dynamically specifying appropriate parameters explicitly in the query. The *Create Geofield* phrase reconstructs a geofield based on a measurement data set and the specification of a generation method. In the next example the user is interested in areas where the concentration of suspended matter falls into two intervals chosen out of the four stated in the query. The distribution is generated during query execution based on chosen: square classification method, continuity model, interpolation, and neighbourhood node selection. SQL *with* clause is used to mark data that may be used later in the query. The answer to the query is displayed as a map (Fig. 2, right).

```
With
 ASMDistr as ( Create Geofield
    From ASMAvg89
    Where square_classifier = 'EstCentreToCorner' and
    continuity_model.id = 12 and
    interpolation.method = 'Kriging' and
    node_search.type = 'Quadrants' and
    node_search.r = 10.25 km and node_search.n = 4 and
    intervals = (110, 165, 220))
Select zone
From ASMDistr
Where  Interval in (0, 3) and
       context_area = mbr(18.38, 49.95, 19.58, 50.55) and
       resolution = 800
```

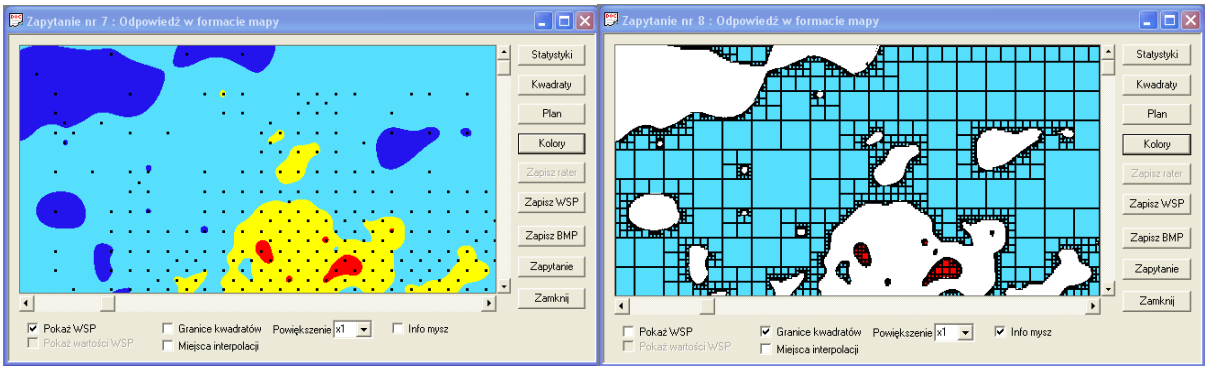

Figure 2. Interval geofield approximation with measurements nodes overlaid (left); the geofield after selection (answer to the example query) with square boundaries (right).

5 SUMMARY

The presented extensions of SQL syntax allow to include in a single compound query: definitions of approximation method for continuous fields, spatial operations on created fields and discrete objects, and conditions on non-spatial data. The use of geofields metadata considerably simplifies the query code. The consistent form of queries makes possible to optimise the query execution process globally on the spatial data server (as opposed to the scripts used in traditional approach).

6 REFERENCES

Bajerski, P, 2002. *Square Classification Methods In The Process Of Spatial Distribution Approximation By Means Of Linear Quadtree*, Technical Report, Silesian University of Technology (in Polish).

Bajerski P., 2000. *Spatial Distribution Server Based on Linear Quadtree*. VII International Conference on Extending Database Technology. PhD Workshop. Konstanz, Germany.

Camara G., Freitas U. M., Casanova M. A., 1995. *Fields and Objects Algebras for GIS* Operations. III Simpósio Brasileiro de Geoprocessamento, São Paulo, Anais, USP.

Camara G., Monteiro A. M. V., Paiva J. A., Gomes J., Velho L., 2000. Towards a Unified Framework for Spatial Data Models. *Journal of the Brazilian Computing Society*, 7(1):2000.

Egenhofer M. J., 1994 Spatial SQL: A Query and Presentation Language. *IEEE Transactions on Knowledge and Data Engineering*, 6(1):86–95, 1994.

Laurini R., Paolino L., Sebillo M., Tortora G.., Vitiello G., 2004. *Dealing with Geographic Continuous Fields – the Way to a Visual GIS Environment*. AVI'04, Gallipoli, Italy. ACM.

Laurini R., Thompson D., 1994. *Understanding GIS*, Academic Press Limited, third printing.

OGIS, 1999. OpenGIS Simple Features Specification For SQL. http://www.opengis.org.

Paolino L., Sebillo M., Tortora G.., Vitiello G., Laurini R., 2003. *Phenomena – A Visual Query Language for Continuous Fields*. GIS'03, New Orleans, USA. ACM.

Rigaux Ph., Schol M., Voisard A., 2002: *Spatial Databases. With Application to GIS*. Academic Press.

Wang F., Sha J., Chen H., Yang Sh., 2000. GeoSQL: *A Spatial Query Language of Object-oriented GIS*. CSIT'2000, Ufa, Russia.