SPATIAL DATA INTEGRATION APPROACH WITH APPLICATIONS IN FACILITY LOCATION

Janis Kampars, Janis Grabis

Riga Technical University, Faculty of Computer Science and Information Technology,
Kalku 1, Riga, Lv-1658, Latvia, kampars@iti.rtu.lv, grabis@iti.rtu.lv

Abstract. Complex decision-making problem, require large amount of data, and data gathering is one of the most complex tasks in the modeling process. Web services and XML are among the most promising technologies for retrieving data from distributed data sources over the Internet. However, different standards are used for providing spatial and non-spatial data, and data retrieval quality of service attributes vary substantially. The objective of this paper is to propose a spatial data integration architecture, which supports data retrieval from heterogeneous, distributed data sources and accounts for quality of service requirements. Application of the integration architecture is demonstrated using a facility location decision-making problem, which requires such spatial data as distances, customer densities and location of competitors. Computational studies are conducted to evaluate data retrieval time and to identify hidden characteristics of different data sources.

Keywords: Spatial data, web services, decision-modeling, facility location

1 Introduction

Complex decision-making problems, for instance in supply chain management, require large amount of data, and data gathering is one of the most complex tasks in decision-modeling. Modern decision support systems often rely on data warehousing to supply necessary data. However, ensuring data quality and timeliness is a challenging task. An alternative to in-house data storage and management is usage of external data sources available over the Internet. This way data can be gathered from the best sources available as needed, and external data providers are in charge of providing appropriate data quality, however immaturity of service registers, poor documentation and frequent changes in service interfaces makes it a complex task. Standardization plays an important role to support data gathering from external often heterogeneous data sources. If data sources are classified as non-spatial data sources and spatial data sources then high level of standardization has been achieved for non-spatial data. XML is used as *franca ligua* for data exchange and web services provide standardized means for data access and processing. XML based data standards are used also in exchange of spatial data [4], [5]. However, none of the available standards has achieved universal acceptance and there is widespread reliance on proprietary technologies. WMS and WFS are spatial data access interfaces for requesting, spatial data and features, respectively [3, 9]. These standards are promoted by the Open Geospatial Consortium (OGC) though they are yet to achieve a wide-spread acceptance. Another OGC standard, which can be used in spatial data exchange is GML [9]. KML, which is an alternative to GML, is a spatial data exchange format actively promoted by Google [4]. Recently, Google has submitted it to the OGC, allowing future harmonization of GML and KML.

On the other hand, importance of spatial data in decision-making has increased what has been greatly facilitated by increasing availability of spatial technologies and data, especially, over the Internet. In order to avoid an increase of data gathering efforts in decision-making associated with data gathering from external spatial data sources, an approach for spatial data integration for decision making purposes is necessary. Lu [6] develops GIS-based platform for intelligent transportation planning. This platform has a strong data presentation layer based on using web mapping services. Luo et al. [7] analyzes Quality of Service (QoS) characteristics of integrated spatial data solutions. However, they do not provide any empirical or experimental QoS evaluation. Stollberg and Zipf [10] demonstrate usage of geoprocessing services in decision support of housing market analysis. They use only spatial data services in their analysis.

The objective of this paper is to propose spatial data integration architecture and to apply gathered data in solving complex decision making problems in supply chain management. The main components of the architecture are spatial data catalogue services, which provide spatial data processing service discovery, QoS data repository, which stores QOS data, spatial data services, and functional spatial data processing adapter, which is the major part of proposed architecture and provides schema mapping, error recovery, QoS evaluation, service selection, local spatial data processing, load balancing and process composition capabilities. Data for decision-making are gathered from both non-spatial and spatial data sources. It is assumed that XML is used as a data exchange format. Decision-modeling data requirements are also specified using XML and mapping between the data requirements and data sources are established. These mappings are used by data retrieval adapter to request data from the external data sources. The adapter is capable to request data from data sources supporting various XML based data exchange standards. It also transforms gathered data in a format suitable for a particular decision-modeling tool. In many situations, similar data are available both in non-spatial and spatial data sources. The advantages of using spatial data sources are possibilities to vary data granularity and use of nearly continuous data. The architecture allows a decision-modeler to retrieve necessary data regardless of standards used at the data source. Application of spatial data integration is demonstrated using a facility location example. The decision-making problem in this example is finding optimal locations for fast food restaurants with regards to customers density, competitors proximity and real estate cost. The necessary data are retrieved from data sources compliant with SOAP, WFS and KLM standards.

The rest of the paper is organized as follows. Section 2 describes the spatial data integration architecture. Application of the spatial data integration in supply chain management and particularly in facility location is demonstrated in Section 3. Section 4 concludes.

2 Integration Architecture

An architecture for spatial data integration is elaborated in this section. The main purpose of this architecture is to ensure that a decision modeling software receives necessary data for solving the selected decision-making problem. Components of the spatial data architecture are shown in Figure 1. The core component is functional spatial data processing adapter (FSDPA). It is referred as to a functional adapter because it does not only provides means for accessing different data sources but also performs data processing multi-step functions needed to prepare data for modeling purposes.

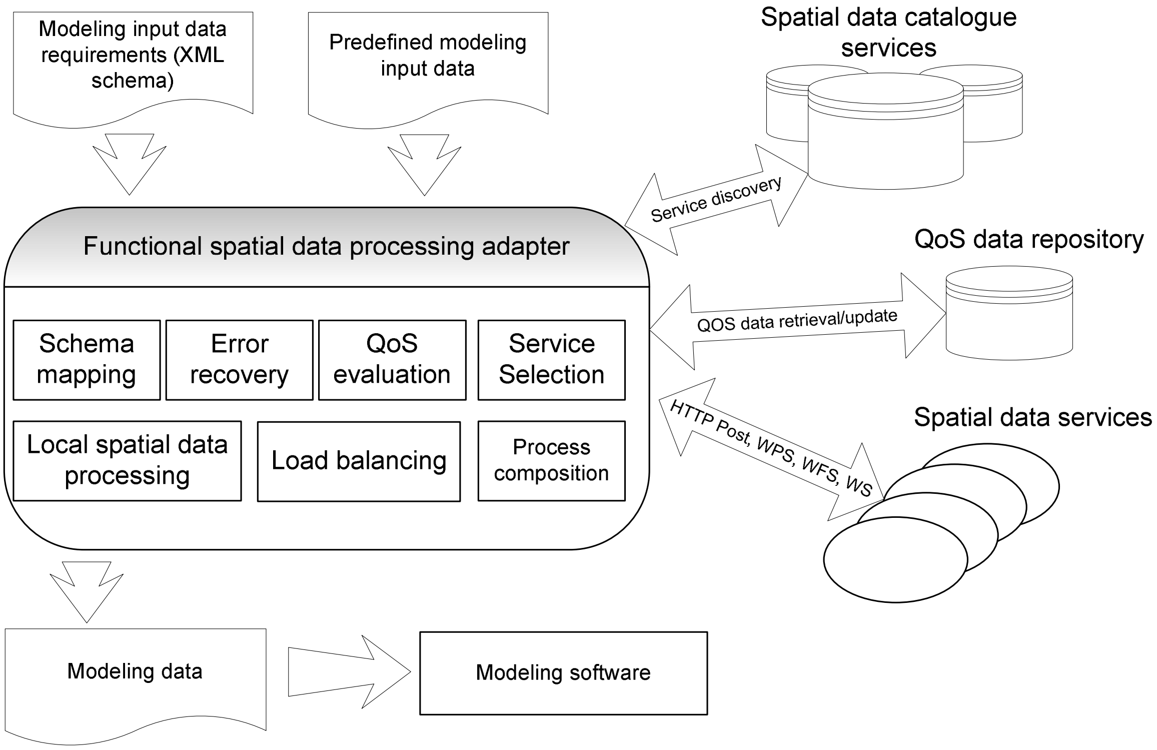
Modeling input data requirements (XML schema)
Predefined modeling input data
Spatial data catalogue services
Service discovery
Functional spatial data processing adapter
QoS data repository
QOS data retrieval/update
Schema mapping
Error recovery
QoS evaluation
Service Selection
Local spatial data processing
Load balancing
Process composition
HTTP Post, WPS, WFS, WS
Spatial data services
Modeling data
Modeling software

Figure 1. Architecture for spatial data integration

FSDPA uses modeling input data requirements defined as an XML schema and predefined modeling input data as input data. The input data requirements are data FSDPA needs to gather from external data sources to satisfy data needs of the modeling software. The predefined modeling input data are data readily available or data necessary for FSDPA. FSDPA retrieves data from external data sources, processes these data and combines them with the predefined modeling input data and provides the combined data to the modeling software in a necessary format. The modeling software is responsible for solving the decision-making problem using the provided input data.

Spatial data and non-spatial data catalogue services are used to discover appropriate external data sources for necessary spatial and non-spatial data, respectively. Services catalogues like Universal Description, Discovery and Integration (UDDI) or OpenGIS Catalogue Services can be searched for suitable services. The Schema Mapping component is used to establish correspondence between modeling input data requirements and available data specifications. XSLT is used for defining schema mappings, and data aggregations and other computational transformations can be applied during the mapping. In cases when there are multiple suitable services the best service is selected using the Service Selection component by analyzing Quality of Service (QoS) data from QoS data repositories (see [7] on selecting spatial services and [2] on selecting non-spatial services). Multiple suitable services allow to provide error recovery. In the case of a service failure, the adapter can switch to the next most appropriate service. If candidate services have limitation like minimum idle time between requests or maximum data requests per day, the load can be balanced between multiple services, thus reducing data retrieval time from APIs and risk of being blocked by certain service provider. In some cases it is more efficient to perform spatial data processing tasks locally rather than requesting processing operations from remote service providers. The Local Spatial Data Processing component provides such functions as some geocoding functions, geographic coordinate conversion and calculation of distances between pairs of points with known coordinates. The Process Composition component is used to define a sequence of data retrieval and processing operations. These operations frequently depend upon each other and therefore the order of their involvement must be specified. The data retrieval process can be defined using general purpose programming languages or WS-BPEL. When the process is executed, QoS data is gathered and sent back to QoS repository, allowing creating more efficient processes in future.

3 Sample Application

Application of the spatial data integration is demonstrated using a facility location decision-modeling problem [8]. The decision-making objective is to determine spatial position of manufacturing or service facilities in order to provide a good customer service and to attain a competitive advantage over competitors. It can be influenced by many different decision-making criteria such as cost, infrastructure, business services, labor, government, customer/market and supplier/resources and competitor related factors [1]. Many of these criteria are associated with spatial measurements. Measurements characterizing number of customers, number of competitors and real-estate cost are used in the particular facility location model proposed in this paper.

The sample application demonstrates: 1) ability of the propose architecture to gather necessary data from heterogeneous data sources; 2) distribution of computational time between data gathering and decision-modeling; and 3) impact of QoS characteristics on data retrieval from distributed data sources.

3.1 Problem Description and Model

The sample facility location problem considered in this paper deals with locating fast food restaurants. There is number of pre-selected potential facility location sites and the total number of facilities to be open is limited. It is aimed to locate restaurants in sites having the largest number of customers and the smallest number of competitors in its proximity and having the lowest real estate costs. Both spatial and non-spatial data are required as problem-solving inputs and the main data sets required are:

- Number of customers in proximity of potential location (spatial data);
- Number of competitors in proximity of potential locations (spatial data);
- Real-estate cost (non-spatial data);
- Distances between potential locations (there is a restriction on the distance between two open facilities) (spatial data).

In order to solve the described facility location problem, a multi-objective mathematical programming model is constructed. The modeling objective is formulated as maximization of an aggregated facility goodness indicator. This aggregated facility goodness indicator is calculated as a weighted sum of several facility goodness indicators corresponding to decision-making criteria (i.e., number of customers, number of competitors and real-estate cost). The model is described in Appendix (more detailed description of the proposed facility location model can be found in [3]. A commercially available optimization software is used to solve the model. It should be noted that the facility location problem computationally is an NP-hard problem, and computational time needed to solve the model directly can be large.

3.2 Data Retrieval Process Composition

In order to solve the facility location problem described in Section 3, three main parameters should be provided for each of alternative locations, namely, customer index a_i, competitors index c_i and real estate cost l_i as well as a distance between potential facilities sites $\Delta_{ii'}$. An XML schema for defining alternative facility locations along with their parameters is developed. An XML document based on this schema is populated with predefined data including list of alternative locations and their addresses.

Spatial and non-spatial data catalogues are used to identify potential data sources and mappings with the data requirements schema are established. Identified data sources and their characteristics are given in Table 1.

The Geocoding service is able to return result in multiple formats, however CSV is used as result contains only two elements – longitude and latitude. The Competitor search engine returns results in two formats from which KML was chosen. Results are divided in multiple KML documents with limited record number in each document. The data retrieval process is relatively fast with small radius as only a few KML documents have to be processed. Use of large radiuses increases response time and can lead to the service overloading. The

Real estate service returns median price per square feet based on city name or zip code. The Population service returns population data in XML format. The results can be filtered by bounding box. Unfortunately the service is unable to filter results based on radius. To overcome this problem, before forming the request to the Population service the local spatial data processing component of FSDPA calculates coordinates of a bounding box that contains the circle in which weighted population is to be calculated. After receiving results from The Population service, the local spatial data processing component filters only those points that fall inside the circle.

The first data source is used for intermediate processing and for calculating $\Delta_{ii'}$. The second source is used to calculate c_i . The third data source is used to calculate l_i . The fourth data source is used to calculate a_i .

Table 1. List of external data sources

#	Data source	Function	Data format	Interface
1	Geocoding service	Converting addresses of facility locations into geographical coordinates	CSV, XML, KML, JSON	Web Service
2	Business directory (competitors) service	Fining spatial location of businesses of specified type	KLM, JSON	Web Service
3	Real estate data service	Fining real estate data for a specified location	XML	REST style Web Service
4	Population data service	Fining number of customers in a specified area	XML	Web Feature Service (WFS)

The mapping between data requirements and available data sources are used to compose the data retrieval process. The retrieval process is described using the UML sequence diagram in Fig. 3. At first coordinates of individual addresses are obtained from The Geocoding web service. Those coordinates are used to query the Competitor search web service and information about nearby competitors is retrieved. The Real estate service is queried by providing city name. At last population data is requested from population WFS. The distance calculation between sites using their coordinates is performed using the local spatial data processing component. The data transformation according to requirements of the decision-modeling software is the last step of data retrieval process. In this case, data are passed to the decision-modeling software as pointers to data arrays in computer memory. This diagram is used to implement the data retrieval process. The process description also could include invoking of error recovery and load balancing components. However, these components were not used in this application.

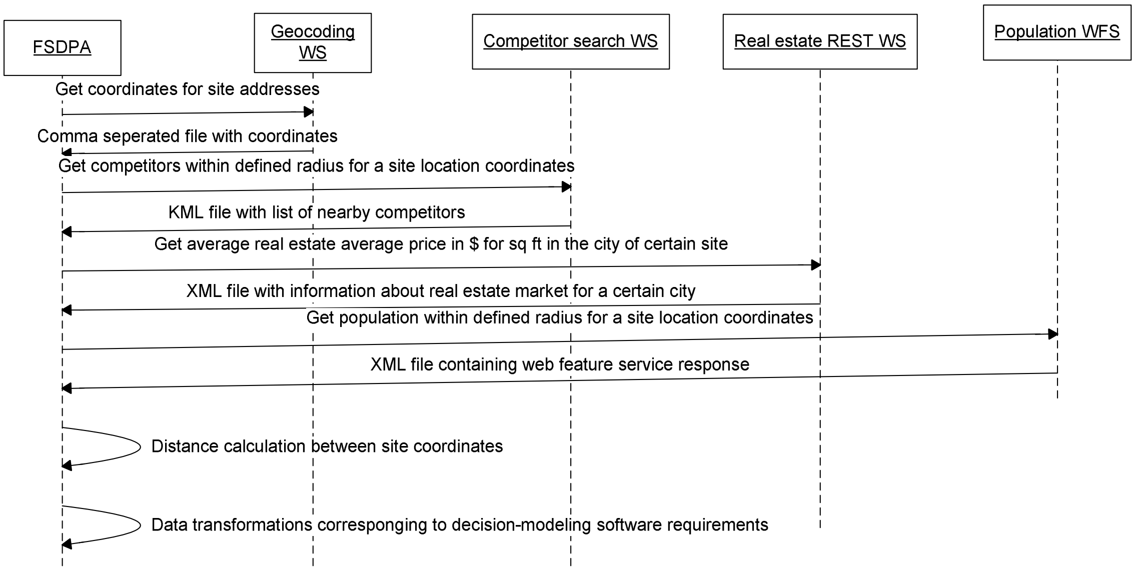
FSDPA
Geocoding WS
Competitor search WS
Real estate REST WS
Population WFS
Get coordinates for site addresses
Comma seperated file with coordinates
Get competitors within defined radius for a site location coordinates
KML file with list of nearby competitors
Get average real estate average price in $ for sq ft in the city of certain site
XML file with information about real estate market for a certain city
Get population within defined radius for a site location coordinates
XML file containing web feature service response
Distance calculation between site coordinates
Data transformations corresponging to decision-modeling software requirements

Figure 3. The sequence diagram of process composition

3.3 Evaluation Time

Computational experiments of data retrieval and decision-modeling are conducted after the data retrieval process has been set-up. Data retrieval and decision-modeling are performed to evaluate computational time and to obtain the final decision-making results. The main parameter affecting data retrieval time is *N*.

Solving time of the facility location model is affected by *N* and *P*. In order to evaluate data retrieval time, data are sequentially requested for 1000 alternative facility locations. The cumulative data retrieval time according to the number of requests made for each data source is given in Figure 4. The Geocoding service is the fastest while the Population data service is the slowest. It should be noted that the Population data service returns a large data set for each request (size of the data set varies) while the Geocoding service returns a few data items. The slope of data retrieval time curve is larger than one for the Competitors and Population services because the cumulative data time is affected by size the return data set and quality of service issues.

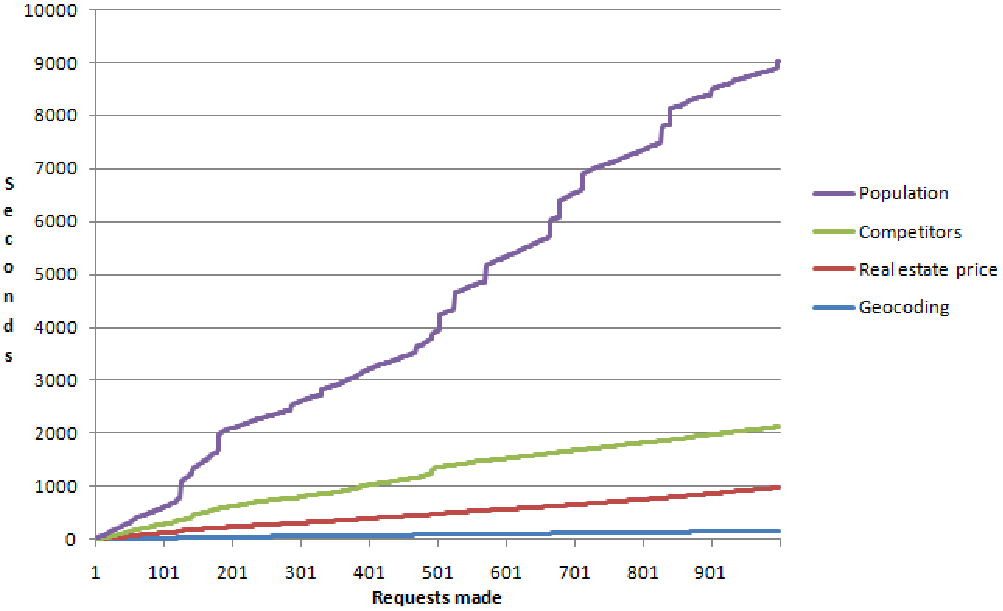

Figure 2. Data retrieval time

Figure 3 compares time spent on data retrieval and time spent on solving the facility location problem. It can be observed that the share of data retrieval time diminishes substantially with increasing *N* because facility location problem is an NP-hard problem while data retrieval time increases linearly. Only the distance calculation in the data retrieval process is N^2. However, it is performed locally and computational time is negligible relative to other operations. Nevertheless, data retrieval time is significant and reaches several hours for large values of *N*.

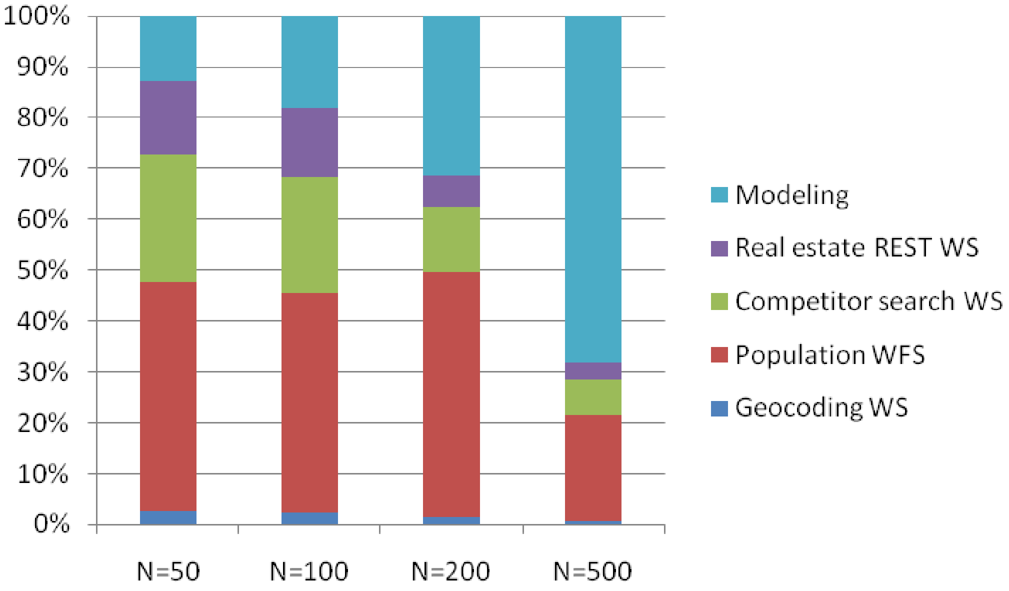

Figure 3. Distribution of evaluation time between data retrieval and decision-modeling

The main emphasis in this paper is on evaluation of data retrieval performance and decision-modeling results are discussed only shortly. Figure 4 shows the total facility goodness indicator for all selected units (a) and the number of open facilities (b) according to *N* and *P*. It can be observed that the facility location model tends to suggest opening the maximum number of allowed facilities. That allows increasing the total facility goodness indicator though marginal returns tend to decrease. If *P* approaches *N* then the sum of open facilities is smaller than the maximum number of allowed facilities because the restriction on minimum distance between open facilities is becoming more important (see Eq. 6 in Appendix). Web mapping services can be used to visualize decision-making results. However, in this case, that was difficult due to the low resolution of population density maps available using WMS.

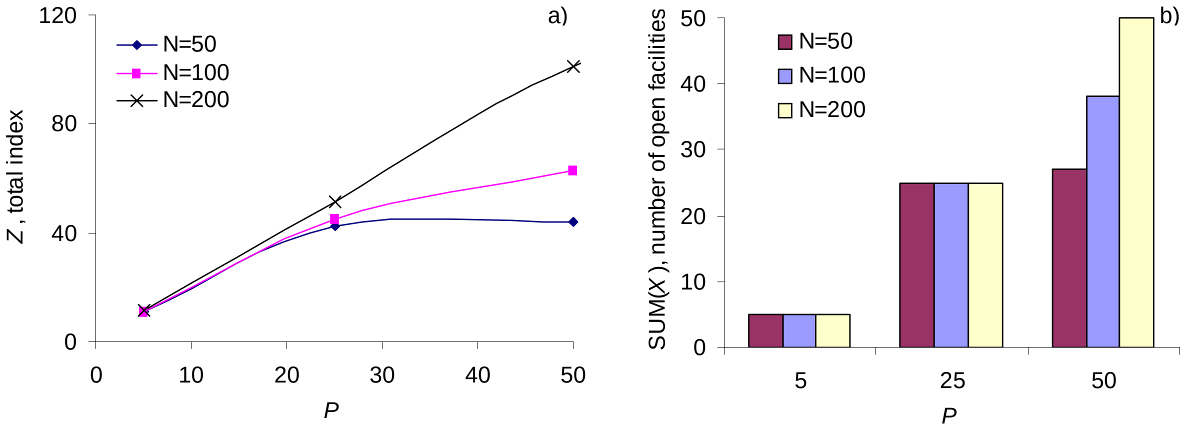

Figure 4. Facility location results

3.4 Quality of Service

Given that data retrieval time constitutes substantial part of the overall facility location evaluation time and services contain missing data, QoS evaluation is an important part of FSDPA. QoS is measured by the response time (i.e. the average time in which service responded to requests made by FSDPA), percentage of dropped request (percentage of requests not processed by service due to service unavailability) and percentage of empty result returned (service queried by the FSDPA returned empty result set due to inadequate granularity). The QoS measures are evaluated over the period of 24 hours, and the results are summarized in Table 2. Data from Table 2 show that QoS varies considerably. The fastest is the Geocoding service. It also has no dropped requests and is able to geocode all of addresses tested. The Real estate service has decent response time and no dropped requests, although only 57.4% of all requests made returned data regarding average real estate prices. The Business catalog service is as stable as two services mentioned above and its response time is satisfactory though it depends upon the coverage radius *r* substantially The dataset quality of this service is average as it returned empty result sets in 15% of all cases. The most unstable was population data service. It dropped about 2% of all requests in periods when it was operating. The response time of this service was long and it returned results only in 46,5% percent of all cases. This service also is the only service, which is not available for prolong periods of times. From the decision-modeling perspective, missing data and low availability are the most undesirable characteristics of data sources.

Table 2. QoS measures

Measure	Geocoding	Real estate	Local business search	Population
Dropped requests %	0	0	0	1,6
Empty result returned %	0	42,5	15,1	53,5
Average response time (seconds)	0,16	0,81	1,14	6,89

4 Conclusion

The architecture for the spatial data integration has been developed in this paper. It also has been applied in gathering data necessary for solving the facility location problem. The main advantages of the proposed architecture are ability to retrieve data from heterogeneous data sources supporting different spatial and non-spatial data distribution standards and process composition. QoS evaluation, service selection, load balancing and error recovery components are also expected to contribute to more efficient data retrieval though formal elaboration of these components is subject to further research. Data gathering from external and, especially, spatial data sources has made possible solving facility location problems, which account for decision-making factors highly important in practical situations. Although advantages of the proposed architecture are more obvious in real-time data processing scenarios, it allowed to reduce cost and data gathering time significantly needed to solve the facility location problem.

The experimental results show that data retrieval time constitutes a substantial part of the problem solving time even though solving the facility location problem is NP-hard problem. The main problem is low availability of some of the services, which makes the data retrieval time unpredictable. The experimental data show that QoS of distributed data sources tend to deteriorate if requests result in large data sets (e.g., in cases of large coverage radius).

The service selection, load balancing and error recovery components depend upon availability of services in service catalogues. Currently, data availability, especially in the public domain, is limited, and catalogues mainly have ad hoc structure.

References

[1] **Bhatnagar R., Sohal A. S.**, Supply chain competitiveness: measuring the impact of location factors, uncertainty and manufacturing practices, Technovation 25, 443–456, 2005.

[2] **Buccafurri c F., De Meoc P.**, **Fugini M., Furnari R., Goy A., Lax G., Lops P., Modafferi S., Pernici B., Redavid D., Semeraro G., Ursino D.**, Analysis of QoS in cooperative services for real time applications, Elsevier Science Publishers, 463-484, 2008

[3] **Chen N., Gong J., Chen Z.**, A High Precision OGC Web Map Service Retrieval Based on Capability Aware Spatial Search Engine State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Springer Berlin / Heidelberg, Volume 4683/2007, 558-567, 2007.

[4] **Du Y., Yu C., Liu J.**, A Study of GIS Development Based on KML and Google Earth, Fifth International Joint Conference , 1581 – 1585, 2009.

[5] **Lu C., Dos Santos R. F. Jr, Sripada L. N., Kou Y.**, Advances in GML for Geospatial Application, Springer Netherlands, 2007.

6] **Lu X.**, GIS-T Web Services: A New Design Model for Developing GIS Customized ITS Application Systems, Springer Berlin / Heidelberg, 875-884, 2006

[7] **Luo Y., Liu X., Wang W., Wang X. and Xu Z.**, QoS Analysis on Web Service Based Spatial Integration Springer Berlin / Heidelberg, 42-49, 2004.

[8] **Owen S. H., Daskin M. S.**, Strategic facility location: A review, European Journal of Operational Research 111, 3, 423-447, 1998.

[9] **Peng Z., Zhang C.**, The roles of geography markup language (GML), scalable vector graphics (SVG), and Web feature service (WFS) specifications in the development of Internet, Springer Berlin / Heidelberg, 2004.

[10] **Stollberg B., Zipf A.**, Geoprocessing Services for Spatial Decision Support in the Domain of Housing Market Analyses - Experiences from Applying the OGC Web Processing Service Interface in Practice, The 11th AGILE 2008 Conference on GI Science (AGILE 2008), 2008.

Appendix

Notation

i - index for potential sites

N - a binary potential sites

P - maximum number of sites to be open

H - number of site selection criteria

a_i - customer index at site *i*

c_i - competitors index at site *i*

l_i - land cost at site *i*

r - coverage radius

B_i^s - set of points *j* falling within *r* around *i*, where *s* indicates a type of point (i.e., customer or competitor)

α_j - number of customers at point *j*

d_{ij} -distance between site *i* and point *j*

$\Delta_{ii'}$ - distance between two potential facilities *i* and *i'*

$v_i = \begin{cases} \exp(-u_{ij}), u \le 1 \\ 0, u_{ij} > 1 \end{cases}$ - weight coefficient, where $u_{ij} = \dfrac{d_{ij}}{r}$

w_k - weight coefficient characterizing importance of each selection criteria *h*, $h = 1,..,H$

X_i - a binary variable indicating if facility is open, $X_i \in \{0,1\}$

Z - aggregated facility goodness indicator

Z_h - facility goodness indicator for selection criteria *h*, $h = 1,..,H$

Model Formulation

The multi-objective function maximizes the aggregated facility goodness indicator

$$Z = \max\sum_{h=1}^{H} w_h Z_h \tag{1}$$

The customer indicator Z_1 is computed as

$$Z_1 = \sum_{i=1}^{N} a_i X_i \ , \tag{2}$$

where $a_i = \sum_{j\in B_i^1} v_i \alpha_j$ is a sum of customers in proximity of potential location *i* exponentially weighted by the distance. In order to calculate the indicator value, a_i are scaled to range between 0 and 1.

The competitor indicator Z_2 is computed as

$$Z_2 = \sum_{i=1}^{N} c_i X_i \ , \tag{3}$$

where $c_i = \sum_{j\in B_i^2} v_{ij}$ is a sum of competitors in proximity of potential location *i* exponentially weighted by the distance. In order to calculate the indicator value, c_i are scaled to range between 0 and 1. Because smaller values of this indicator are preferable, scaling is performed using $c_i^{'} = 1 - \frac{c_i}{\max\limits_{\forall i}(c_i)}$.

The real estate cost indicator Z_3 is computed as

$$Z_3 = \sum_{i=1}^{N} l_i X_i \ , \tag{4}$$

where l_i is scaled to range between 0 and 1 giving the preference to smaller values similarly as for Z_2 .

The maximization is performed subject to the following constraints. The constraint (5) restricts the number of facilities to be open:

$$\sum_{i=1}^{N} Y_i \le P \ . \tag{5}$$

The constraint (g) implies that the distance between two open facilities should larger than *r*:

$$\Delta_{ij} X_i X_j \le r, \forall i, j, i \ne j \ , \tag{6}$$

In the mathematical model, this constraint is transformed using proxy variables to eliminate non-linearity.

In order to compute the aggregated location goodness indicator, a weighted sum of individual facility goodness indicators is computed. Importance of each indicator can be determined according to the results of empirical studies on practical importance of different facility location criteria. The model presented above uses just three facility location criteria although other criteria can be incorporated if necessary.