Transactions in GIS

Research Article Transactions in GIS, 2015, 19(1): 107–128

Integrating Spatial Data Linkage and Analysis Services in a Geoportal for China Urban Research

Xinyan Zhu,* Bing She,* Wei Guo,* Shuming Bao† and Di Chen*

*LIESMARS, Wuhan University
†China Data Center, University of Michigan

Abstract
Many geoportals are now evolving into online analytical environments, where large amounts of data and various analysis methods are integrated. These spatiotemporal data are often distributed in different databases and exist in heterogeneous forms, even when they refer to the same geospatial entities. Besides, existing open standards lack sufficient expression of the attribute semantics. Client applications or other services thus have to deal with unrelated preprocessing tasks, such as data transformation and attribute annotation, leading to potential inconsistencies. Furthermore, to build informative interfaces that guide users to quickly understand the analysis methods, an analysis service needs to explicitly model the method parameters, which are often interrelated and have rich auxiliary information. This work presents the design of the spatial data linkage and analysis services in a geoportal for China urban research. The spatial data linkage service aggregates multisource heterogeneous data into linked layers with flexible attribute mapping, providing client applications and services with a unified access as if querying a big table. The spatial analysis service incorporates parameter hierarchy and grouping by extending the standard WPS service, and data-dependent validation in computation components. This platform can help researchers efficiently explore and analyze spatiotemporal data online.

1 Introduction

Enabling faster scientific research capacities is a central theme guiding e-Science/e-Research investigations (Hey and Trefethen 2005). In the geospatial community, developing specifications to encapsulate geospatial data and geoprocessing functions as standard web services is also an active area of study. The Open Geospatial Consortium (OGC) compliant services are probably the most widely known standards specifications, such as the Web Map Service (WMS), Web Feature Service (WFS), and Web Processing Service (WPS) (Yue et al. 2010). Effectively integrating these data and analytic services across multiple domains are critical tasks in the era of spatial Cyberinfrastructures or CyberGIS (Yang et al. 2010; Wang et al. 2012). This is particularly true for urban research, since socioeconomic processes are inherently multi-dimensional.

Studies of cities, metropolitan areas, and urbanized regions increasingly rely on multisource heterogeneous data, as well as analytic models to support interactive exploration

Address for correspondence: Wei Guo, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan Hubei 430079, China. E-mail: guowei98032@gmail.com

Acknowledgements: The research is funded by the National Science and Technology Pillar Program (2012BAH35B03), the Wuhan Technology Program (2013010501010146), the Fundamental Research Funds for the Central Universities and Open Research Fund of LIESMARS. We greatly appreciate the constructive and valuable comments from the anonymous reviewers and editor. We would also like to thank Edward C. Chao at Data Numerica Institute for providing the StatGL library and Myunghwa Hwang at Arizona State University for providing the Python scripts for linking the PySAL library.

doi: 10.1111/tgis.12084

and modeling processes (Pettit et al. 2012). Although more data are becoming publicly available through open standards and web APIs, ad hoc data access and management are still the prevalent approaches in urban research (Sinnott et al. 2011). Due to the growing need for unified data access, many geoportals were developed to integrate distributed data sources and spatial analytic tools. These geoportals create online environments for studying a wide array of issues, including regional development, energy consumption, and human-environment interaction.

A Geoportal for urban research typically integrates a number of services such as report generation, thematic mapping, and spatial analyses. These services all require access to the underlying data sources (Tomko et al. 2012). Data sources include the OGC services (most commonly WFS), as well as locally held databases that contain non-spatial or time-series attributes. Consequently, data sources that refer to the same geospatial entity might be located in multiple instances of OGC-compliant servers or databases. Researchers must synthesize these data sources to meet analysis requirements by connecting the data and possibly generating new data (Andrienko and Andrienko 2006). In addition, different data sources have varying capacities to represent attribute semantics that contains its name, unit, data type, and other meta-information. For example, open standards like WFS still lack sufficient expression for the attribute semantics (Zhang et al. 2010). Data linking and annotation tasks are thus often left to client applications or other services, which are unrelated to their core functionalities and might lead to potential inconsistencies. Therefore, geoportal developers need to design a dedicated service to add, match, transform, and annotate multisource heterogeneous data into several linked layers. Such a service provides a unified access to client applications and services as if they were operating on a big table.

Analysis of urban data requires various methods that combine complex geocomputation and interactive visualization. Complex spatial analysis methods contain many interrelated parameters with their own presentation and validation semantics. Thus, users often go through a time-consuming trial and error process to figure out how the system works. The OGC's WPS standard is a valuable starting point for building an informative interface. Research in the scientific workflow community that tackles parameter validation in complex data analytics can also be adapted to the characteristics of spatiotemporal data (Kumar et al. 2010; Gil et al. 2011). A user-centric online workbench can therefore be created – with a reasonable decomposition of computation and visualization components – to provide better presentation, validation, and understanding of complex spatial analysis methods.

This article presents the design and implementation of a spatial data linkage and analysis service integrated in a web-based platform for China urban research. The spatial data linkage service serves as a middleware tier that re-organizes heterogeneous multisource data into several linked layers. Each linked layer consists of virtual attributes that map into physical attributes or their combinations, with auxiliary semantics defined. The spatial analysis service provides enriched metadata for the parameters required for spatial analytic methods to support dynamic form display and input validation. On this platform, researchers can efficiently explore and analyze large amounts of spatiotemporal data directly.

The remainder of the article is organized as follows. Section 2 presents related work. Our background and the conceptual design are outlined in Section 3. Sections 4 and 5 address the mechanism and architecture for the spatial data linkage and analysis services. Section 6 describes platform implementation. Section 7 gives a case study as well as the discussion. The conclusion is provided in Section 8.

2 Related Work

There was an array of initial visions for publishing geospatial information over the Internet. These approaches included building a geolibrary to allow users to search map resources by places (Goodchild et al. 1999), and the earth system science workbench for archiving and publishing research data (Frew and Bose, 2001). Pioneering work such as the Alexandria library project built large collections of distributed map resources (Frew et al. 2000). Geoportals have since proliferated, and standard metadata and web services are used to ensure interoperability (Amirian et al. 2010). For example, Fry et al. (2012) built a geoportal for managing socio-economic data for Wales. This system uses standard metadata such as the Dublin Core and services such as WMS for spatial data display. In many cases, however, geoportal administrators still face a challenge when integrating data services of various kinds. Tomko et al. (2012) described the development of an e-Infrastructure capable of integrating related urban data sources published through WFS services, database endpoints, and also social media APIs. Data integration is often performed in the client side. For example, Li et al. (2011) proposed an AJAX-based multi-catalogue search solution to better facilitate discovery of separate OGC Web services into an unified interface; Ho et al. (2012) developed a client-side visual analytics framework for the flexible combination of multi-source data.

Data source level metadata provide the basic data characteristics including provider information and collection techniques. Data analytics in urban research though, requires semantic information down to the attribute level. Semantic technologies are often applied in such cases to explicitly model the object relations through ontologies (Fonseca et al. 2002; Stadler et al. 2012). Researchers have investigated techniques such as the resource description framework (RDF) for feature level annotation (Batcheller and Reitsma 2010). The OGC has also published standards for querying geospatial semantic data (OGC 2012a). Defining formal ontologies requires well-established vocabularies and concepts. Therefore, in some recently developed geoportals where large amounts of attributes are increasingly added, such as AURIN (Tomko et al. 2012), the basic semantics for attributes were defined in metadata to assist data analytics. This work focuses on the flexible management and combination of multisource heterogeneous data at the service level, and provides a lightweight solution to define basic attribute semantics for data analytic needs.

Built on data services, spatial analysis services provide a wide range of interactive visualization tools through standardized communication protocols. The GeoVISTA studio creates probably the first complete workbench environment that integrates diverse spatial analysis and geovisualization components (Takatsuka and Gahegan 2002). Anselin et al. (2004) demonstrates an early effort to put exploratory spatial data analysis tools online. A vibrant research agenda termed Geovisual Analytics (Andrienko et al. 2010) has now emerged, and is actively investigated in diverse fields such as ecology (Auer et al. 2011), environmental planning (Ghaemi et al. 2009), and public health (MacEachren et al. 2008). Researchers in the domain of scientific workflow systems have extensively studied parameter semantics and validation issues in complex data analytic tasks (Berkley et al. 2005; Deelman et al. 2009). The parameter space of the Wings workflow system, for example, is constrained by the workflow components and the input data (Gil et al. 2010, 2011). OGC also put a lot of effort in defining the WPS service standard that regulates the input data of a given method (Schut and Whiteside, 2007). Broadly speaking, the attribute and parameter semantics can be incorporated into the data provenance (Yue et al. 2011) and model provenance (Anselin and Rey, 2012) architecture. Our work incorporates parameter validation and its meta-information into the OGC's WPS standard, and supports data-dependent validation in computation components.

3 Motivation and Objectives

3.1 Web-based Data Integration and Analysis Platform

This work is built on an online environment for integrating and analyzing heterogeneous multi-source data we developed previously (She et al. 2012). In this environment, users from different backgrounds, including academia, government, or the general public, can browse data, make reports, and create thematic maps on demand. Over the last few years, spatiotemporal data from diverse fields were gradually integrated, including historical population and economic censuses, yearly government statistics, and environmental data observations. Researchers have begun to use open source spatial analysis software more frequently (Rey 2009). We have integrated several frequently-used analysis methods into our platform; PySAL and R are mainly used. PySAL is an open-source spatial analysis library that integrates a growing number of spatial analytical functions (Rey and Anselin 2010). R is a dynamic language that contains general statistical functions as well as packages for spatial analysis (R Development Core Team 2011). The Java topology suite (JTS) is also used extensively for basic spatial operations (Vivid Solutions 2013).

3.2 Need for Spatial Data Linkage

The data for each urban district is distributed across multiple tables, or even databases. For example, the data for each city in China, which include age composition, employment status and environment condition, are stored separately in the population census, the economic census, and the government statistics. Some of these data sources contain a spatial attribute and are published through WFS, while others do not. Users often need to put these multi-source data in a single report, or construct a new attribute by aggregating existing attributes during an analysis. Moving all these data into the same database is technically infeasible due to both issues of data update and heterogeneity of attribute representation. It can also be politically challenging.

These attributes from different data sources tend to have varying degrees of expressive capacity describing their unit, level of measurement, and other semantic information. Attribute semantics define how attributes should be represented, transformed, and calculated. For example, to display an attribute value properly in a report, we need its name, unit, and data source information. If it is numerical, the number of decimal places is required. In data analytics, the primitive data type and level of measurement of an attribute is needed for parameter validation. Therefore, we need a unified data access entrance to query heterogeneous multi-source data that manages the processes of request decomposition, data transformation, response assembly, and adding basic semantic information.

3.3 Making Sense of Spatial Analysis Methods

Geovisual analytic platforms assist researchers in gaining insights by interactively and iteratively exploring data through various methods with different parameters. The construction of an informative user interface is critical to help researchers understand and operate these methods. Therefore, our platform explicitly incorporates parameter semantics and validation in the web Application Programming Interfaces (APIs). Parameter semantics include the default value, data types, validation rules, and descriptions, as well as how parameters are correlated and grouped. The data input of the *DescribeProcess* operation in a WPS service includes the assignment of default values and allowed values (Schut and Whiteside 2007). Our service extends this operation to allow a hierarchy of inputs representing parameter groups, as well as a composite parameter that allows a dynamic number of parameters. The grouping and switching of parameters together represent the parameter hierarchy and interrelation for this particular method.

Data-dependent validation is necessary since the default value and valid range of parameters are often dependent on the specific spatial data selected. For example, the calculation of LISA statistics requires the construction of a spatial weight matrix first. If the matrix has all its elements equal to zero (possibly due to an inappropriate threshold value set by the user), the subsequent computations would be meaningless. In this case, the valid range of the threshold should be recalculated to have a minimum value of the shortest distance among all the spatial object pairs.

3.4 Conceptual Design

The spatial data linkage and analysis services work together to allow users to more efficiently synthesize and analyze data. Figure 1 shows the basic composition of different components in the platform. A linked layer provides middle-tier virtual access to underlying data sources. The original attributes or their aggregation in the underlying data sources are mapped into virtual attributes in the linked layer, incorporating basic semantics. The spatial data linkage service supports flexible management of linked layers, including the definition, mapping, organization,

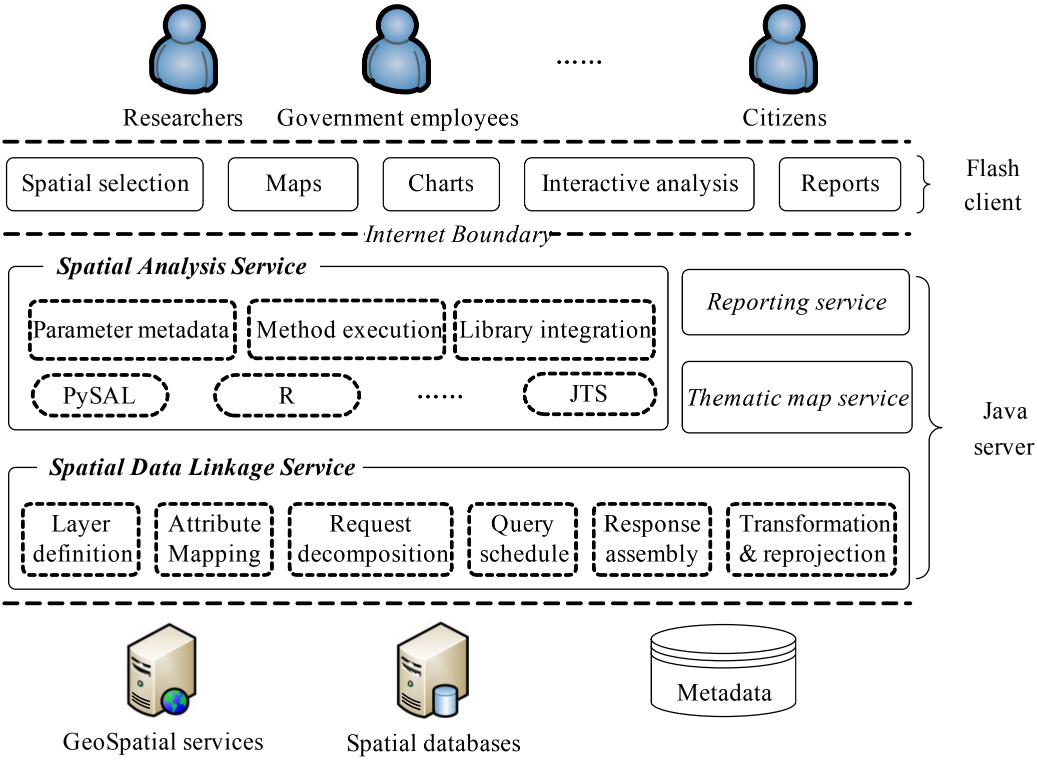


Figure 1 The component composition of the platform

and linkage of data sources and virtual attributes. These operations are provided to Geoportal administrators to construct specific linked layers according to their requirements. When the Geoportal is deployed, other services and client-side applications can query the data from linked layers through unified access APIs.

Spatial analysis of vector data can be both exploratory and confirmatory, and the exploratory phase often requires interactive map and chart displays to help users better understand the trends and patterns. Therefore, we define an analysis method as a composition of a computation component and a visualization component. The method metadata is thus a collection of parameters organized in a certain structure that defines the input for the respective computation and visualization components. The method definition defines the hierarchy and categorization of the parameters and methods. The API directly extends the inputs of the WPS standard. This approach prevents client applications from hard-coding this information. A server-side validation API is defined in the computation component to deal with data-dependent validations. Sections 4 and 5 discuss the design and features of the two services in detail.

4 Spatial Data Linkage Service

The process of defining the linkage service consists of two parts: defining the data sources with their virtual attributes, and constructing linked layers that consists of multiple sources connected by linking attributes. Administrators can define the data sources, the virtual attributes, and the linked layers through a set of RESTFul web APIs of the spatial data linkage service. During the data query phase, the spatial data linkage service will handle the jobs of request decomposition, data transformation, and response assembly.

4.1 Data Source Definition and Attribute Mapping

A linked layer can be seen as an aggregate of multiple data sources. Data sources in a linked layer can be one of the following two types: (1) spatial data sources published through a Web Feature Service (WFS); or (2) attribute data sources stored in a relational database. The spatial data sources represent WFS feature types published in OGC-compliant servers, designated by the server address and the name of the feature type. Our current design only targets the Simple Features that serve the general purpose, instead of particular scenarios in Geography Markup Language (GML) application schemas, which consist of nested structures for complex objects (OGC 2012b). A Spatial Reference System Identifier (SRID) is assigned to the spatial data source upon definition. Attribute data sources represent the non-spatial attributes directly connected to a relational database, identified by the database identifier and the table name. Attribute data sources are further divided into basic and time-series data sources. Basic attribute data sources contain non-spatiotemporal attributes, while the time-series attribute data source stores temporal attribute data at various temporal scales. The temporal scale reflects the update or collection frequency of the data. Six types of temporal scales are defined in our current design: yearly, monthly, daily, hourly, customized interval, and continuous. In addition, a linked layer can itself be a special kind of data source, called a reference layer. This is helpful when the query requires information from two different layers such as: "how many hydrological stations are there in this city?".

These data sources collectively provide a large pool of physical attributes for the linked layers. Despite attribute semantics, many attributes have to be transformed or combined for later data queries, including temporal attributes that require multiple fields in the database, or

Table 1 The characteristics of five main meta-attributes

Name	Permitted values	Default value
Unit	kilometers, Yuan, million persons . . .	empty
Data type	integer, double, percentage, permillage, string, geometry, . . .	inferred from the original type in the data source
Level of measurement	nominal, ordinal, interval, ratio	depending on data type
Decimal places	>0	0 if data type is integer, 2 otherwise
Display names	multilingual names	attribute code
Description	string	empty

attributes in a reference layer that need to be recalculated. Therefore, we designed a virtual attribute to be a consistent representation that mapped physical attributes from heterogeneous forms.

A virtual attribute can be of four types: basic, spatial, temporal or composite. Except for the composite attribute, all other types of virtual attributes map to a single physical attribute without transformation, we thus call them primitive virtual attributes. The mapping is designated by the identifier of the data source and the attribute code of the physical attribute. The identifier of the data source acts as a namespace, preventing potential naming conflicts. The basic virtual attribute commonly maps to physical attribute in a basic attribute data source, but can also map to non-spatial attributes in a WFS data source. The spatial virtual attribute maps to the geometry attribute in a WFS data source, while the temporal virtual attribute maps to the attributes in a temporal attribute data source. A composite virtual attribute corresponds to an aggregation of other types of virtual attributes. In most cases, the mapping corresponds to basic arithmetic computation, but it can also involve spatial computation such as calculating the centroids of polygons. The attributes in reference layers always need transformation into composite virtual attributes through customized procedures. These computation routines are incorporated as plug-ins in the spatial data linkage service.

The attribute semantics in the current design consists of a set of meta-attributes as shown in Table 1. The default values are useful when virtual attributes are automatically generated, and could be adjusted by administrators afterwards. A unit is defined through the spatial data linkage service, including the label as well as the transformation rules between different units. These defined transformation rules are used in data query processes. The level of measurement follows Steven's categorization (Stevens 1946) and is useful such as when filtering attributes in an analytic API. Data types define the data structures of the attribute which, together with the number of decimal places, decides how to display numbers in a report or in a map legend. Currently the number of decimal places in Table 1 applies only to the non-spatial attributes.

Each of the data sources and virtual attributes owns a global identifier to support sharing among different linked layers. However, virtual attributes are still tied to particular data sources if the global identifiers of the data source are used in the attribute definition. For example, our platform has four separate data sources representing the 2004 economic censuses for the different administrative levels, but all these data sources have essentially the same physical attributes. To avoid defining the virtual attribute four times, a local identifier (non-unique) is assigned to each of the four data sources used in the linked layer. In this way, a virtual attribute can be associated with a group of data sources. When a data source is defined, a set of virtual attributes will be automatically generated corresponding to the physical attributes (except the composite type which requires transformation). Alternatively, they can be constructed by Geoportal administrators through the web APIs.

4.2 Linked Layer Construction

The construction of a linked layer is a three-step process. The first step is to bind a set of data sources, and attach virtual attributes. Two data sources with the same local identifiers cannot be bound to the same linked layer, preventing ambiguous data source references in the virtual attributes. By default, all the automatically generated virtual attributes corresponding to the bound data sources are attached. These attached virtual attributes are reorganized into a set of themes. Themes are groups of attributes, which serve to display virtual attributes to end users through a meaningful structure. Themes can also be nested to show a particular hierarchy. Optionally, a subset of virtual attributes can be selected to represent meta-information about the queried data items, such as the name or area of a province. Such attributes are called descriptive virtual attributes.

The second step is defining the linkage strategy. First, a basic virtual attribute needs to be set as the default identifier attribute in the linked layer, named as DIA for brevity. The DIA will correspond to a data source in the linked layer, which now becomes the default data source. Then, a matching virtual attribute will be assigned to each of the bound data sources except for the default one. The matching virtual attribute can either be basic or composite. The basic virtual attribute is used when the code values are consistent with the DIA, while the composite virtual attribute is used when the data source has different coding schemas from the DIA. A common example is the different naming conventions in coding administrative units regarding prefixes or postfixes. The matching composite attribute will be defined using a computation plug-in as described in Section 1.1. Such a plug-in either performs the matching at runtime (such as removing the leading zeros), or reads the matching results pre-generated externally (such as using a semantic harmonization tool) when the matching phase is complex or time-consuming.

The third step is to define a set of meta-attributes for the linked layer including its name, a query limit that defines the maximum number of features that can be returned, and a default geometry attribute used in spatial computation (selected from the spatial virtual attributes). After this three-step process, a global identifier will be assigned to the linked layer, and the layer will be added to the layer store.

4.3 Architecture

The spatial data linkage service provides two categories of operations. The first category is for management of linked layers, used by Geoportal administrators. The second category is for data query of linked layers, used by other services and client-side applications. Figure 2 depicts the high-level architecture of the spatial data linkage service. *VAttribute* stands for virtual attribute. The management operations are handled by a set of management classes, including the creation, modification and storage of data sources, virtual attributes, and the linked layers. Except for some built-in unit types, new units as well as the transformation rules for existing units can be defined through the management APIs. These operations are currently exposed to Geoportal administrators through a set of RESTful Web APIs. When a data source is added, a set of default *VAttributes* are generated by sending appropriate capacity requests to the

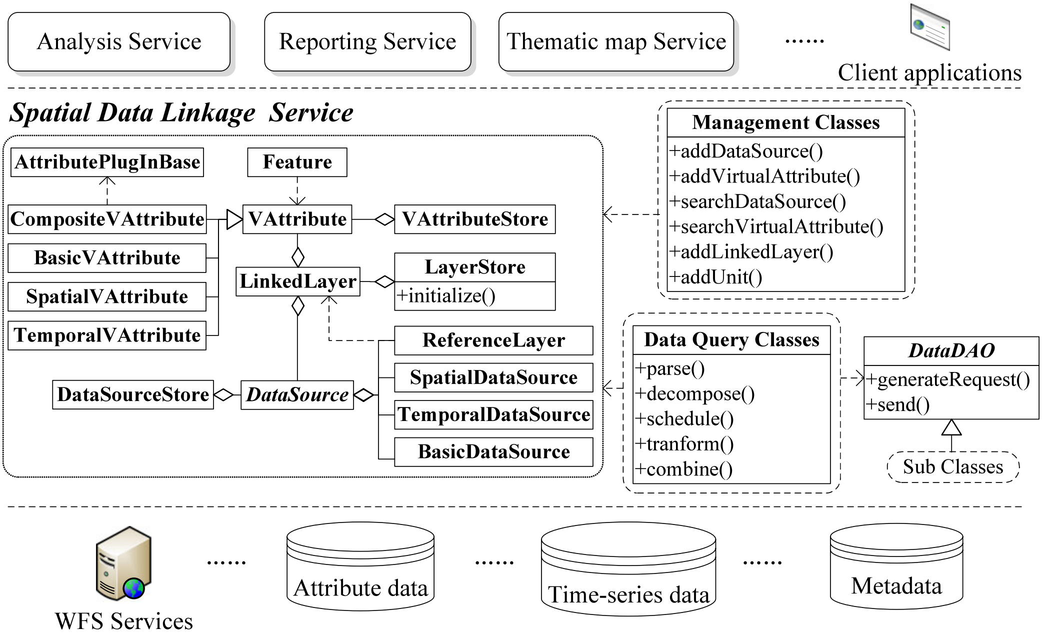


Figure 2 The high-level architecture of the spatial data linkage service

corresponding data storage. For example, a *GetCapabilities* request is sent to a map server publishing a WFS service for a spatial data source. Composite attributes are created by either defining an arithmetic expression inside the request or, alternatively, uploading a plug-in extending the *AttributePlugInBase* that incorporates routines to deal with transformations. The final definitions are stored as XML files in a server directory assigned to the spatial data linkage service. A set of search operations are defined to retrieve the metadata of the data sources and virtual attributes to assist Geoportal administrators to construct the linked layers. Upon initialization, the *LayerStore* will initialize the store of linked layers by parsing the definition files.

There are three types of data query operations for querying the metadata of linked layers and the actual data. The first one is *GetCapabilities* which retrieves the list of all linked layers (returning only its meta-attributes). The second is *DescribeFeatureType* which retrieves the virtual attributes of a linked layer with its meta-attributes. The client applications can get the virtual attribute information as well as the data source information. The third operation is *GetFeature,* compatible with WFS's *GetFeature* operation, which is handled by a set of query classes as shown in Figure 2. This query process consists of following three phases:

1. **Request parsing:** The linked layer and the required virtual attributes are first identified through their keys in the request. The request is compatible with the WFS specification. These virtual attributes are separated into groups by data sources. Before separation, composite virtual attributes are decomposed into a set of primitive virtual attributes. A set of atomic requests will then be constructed for each attribute group, which can either be a WFS request or a database SQL command. Since the WFS specification allows multiple data queries in the same request, atomic requests directed to the same OGC server are combined into one request. After this grouping and combining process, there are now a

number of sub-requests. Minor changes were made to the WFS specification to allow unit transformation on the server side. For example, the total population attribute is stored as *persons*, but the request can specifically require the unit to be *million persons*.

2. **Data retrieval:** If the original request uses the values of the DIA to identify spatial objects, the matching attributes of the accessed data source are used in the query construction for respective sub-requests. These sub-requests are then sent using respective access strategies. If the original request uses a spatial predicate to identify spatial objects, the sub-request that contains this spatial predicate will be sent first (a WFS request). When the values of the DIA are returned from this particular response, other sub-requests will be constructed and sent out.
3. **Response assembly:** Responses are first transformed into *Feature* objects, as shown in Figure 2, using the respective parsers. Composite attributes are calculated from the decomposed attributes through the composite attribute computation logic. Spatial re-projection is conducted if the projection of the spatial data source is different from the requested projection. For those attributes that have requested units different from their original units, transformation is done based on the associated rule. Finally, the responses are assembled for client applications and services. The response data are wrapped into the format of the Geography Markup Language (GML) and transmitted through HTTP protocols. Minor changes were made to the WFS specification to allow for the temporal attribute query and expression. Future work will investigate other formats with more expressive power such as WaterML 2.0 (Valentine et al. 2012), an application schema for GML 3.2.1.

5 The Spatial Analysis Service

Similar to the spatial data linkage service, the process of defining the method data is configured by administrators through the RESTFul web APIs provided by the spatial analysis service. When users are conducting certain analysis, the spatial analysis service, together with the front-end client, will deal with the process of parameter validation, environmental setup, method invocation, and result display.

5.1 Defining Method Metadata

The WPS standard already includes the *LiteralData* type that represents primitive data types such as *int* or *string* (Schut and Whiteside 2007). The client-side will validate the input according to the data type and the allowed values. Three type extensions of input parameters are defined. The first two correspond to one or multiple virtual attributes used by the spatial linkage service; users can also construct a composite virtual attribute represented by an arithmetic expression on the client-side. The third represents a composite parameter that has a set of options, each consisting of a separate subset of parameters. This design leads to a dynamic form on the client-side. When users change the option, the form will display a different subset of parameters from those previously displayed. The WPS standard has provided the *ComplexData* input type that supports customized data structures. Instead, this work defined these extensions directly at the *DataInputs* level since the entire representation schema was changed in this service. Currently, the analytic outputs are wrapped with the customized data structures inside the *ComplexData* due to the high variability in result representations from different analytic methods.

A method definition consists of a number of parameters, organized into a set of groups for clarity. Table 2 shows an example definition of the Local Moran's I method (Anselin 1995), represented as a process description in the *DescribeProcess* operation of a WPS standard service. The method contains three sets of parameters, one for the attribute, one for the spatial weight matrix, and the other for hypothesis testing. Each parameter has its meaning and permitted values. The *AttributeInput* and *CompositeInput* represent the attribute parameter and composite parameter. The *spatialweight* parameter is an example of a composite parameter consisting of two parameters representing weight type and standardization method. The *lom* attribute in the *AttributeInput* signifies a virtual attribute at a *ratio* level of measurement. The *show* attribute inside the *<InputGroup>* node defines whether this parameter group will be initially shown (such as folded in the user interface). Similar to parameter groups, the methods are divided into a set of method groups, which can also be nested to display a certain hierarchy. A method group has a *category* attribute of two possible types: *spatial* and *spatiotemporal*, useful in filtering the virtual attributes of a linked layer displayed in an attribute parameter.

5.2 Data-dependent Parameter Validation

The method metadata provides instant client-side validation, but it is limited to predefined values. To incorporate data-dependent parameter validation at runtime, the spatial analysis service embeds a parameter validation phase in the computation component that decides the parameter validity based on the spatial characteristics of selected data. The validation is triggered when the data is retrieved and parameter values are parsed. If the validation fails, a readable context-dependent message is constructed and returned, along with the new validation rule that overwrites the existing rule that was read statically from the method metadata. These two validation strategies, provided separately by the metadata and computation components at runtime, can be seen as corresponding to component and data constraints respectively, in a scientific workflow system.

5.3 Method Composition

Both the computation and visualization components are published through the spatial analysis service, but the execution mechanisms are different. The computation components execute on the server-side by linking to different libraries and tools for spatial analysis, while the visualization components exist as pre-complied resources and are retrieved and activated on the client-side after receiving the output from the corresponding computation components. The separation of computation and visualization is not absolute. A computation component may perform no real analysis but simply retrieve the data from the spatial data linkage service; and a visualization component might include computation routines, such as numerical summarization for a selected sub-dataset.

5.4 Architecture

Figure 3 depicts the high-level architecture of the spatial analysis service. The method repository reads the method definitions at system startup, and instantiates the computation components corresponding to the requests. The execution relies on appropriate environment setup, since a hybrid approach was adopted by integrating different analysis libraries and tools. The cache function is useful when further requests rely on previous calculations such as a re-run of a random simulation test.

Table 2 Part of the Global Moran's I method defition respresnted by the modified process description in the *DescribeProcess* operation. The bolded nodes represent the extensions to the standard

```
<ProcessDescription wps:processVersion="1" storeSupported="True" statusSupported="True">
 <ows:Identifier>SGMI</ows:Identifier>
 <ows:Title>Single Moran's I(Single Variable)</ows:Title>
 <DataInputs>
  <InputGroup>
   <ows:Title>Basic Paramters</ows:Title>
   <AttributeInput minOccurs="1" maxOccurs="1">
    <ows:Identifier>x</ows:Identifier>
    <ows:Title>Varible X</ows:Title>
    <lom>ratio</lom>
   </AttributeInput>
  </InputGroup>
  <InputGroup>
   <ows:Title>Spatial Weight</ows:Title>
   <CompositeInput minOccurs="1" maxOccurs="1">
    <ows:Identifier>spwgt</ows:Identifier>
    <ows:DataType ows:reference="xs:string">string</ows:DataType>
    <Options>
     <Option name="Rook contiguity">
      <Input minOccurs="0" maxOccurs="1">
       <ows:Identifier>order</ows:Identifier>
       <ows:Title>Order</ows:Title>
       <ows:Abstract>. . .</ows:Abstract>
       <LiteralData>
        <ows:DataType ows:reference="http://www.w3.org/TR/xmlschema-2/#integer">integer</ows:DataType>
        <ows:AllowedValues>
         <ows:Range>
          <ows:MinimumValue>1</ows:MinimumValue>
          <ows:MaximumValue>10</ows:MaximumValue>
         </ows:Range>
        </ows:AllowedValues>
        <DefaultValue>1</DefaultValue>
       </LiteralData>
      </Input>
      <Input minOccurs="0" maxOccurs="1">
       <ows:Identifier>includeLowerOrders</ows:Identifier>
       <ows:Title>Include lower orders</ows:Title>
       <LiteralData>
        <ows:DataType ows:reference="http://www.w3.org/TR/xmlschema-2/#boolean">boolean</ows:DataType>
        <ows:AnyValue/>
       </LiteralData>
      </Input>
      . . . . . . .
     </Option>
     . . .. . .
    </Options>
   </CompositeInput>
   <CompositeInput minOccurs="1" maxOccurs="1">
    <ows:Identifier>standardization</ows:Identifier>
    . . . . . . .
   </CompositeInput>
  </InputGroup>
  <InputGroup>
   <ows:Title>Test</ows:Title>
   . . . . . . .
  </InputGroup>
 </DataInputs>
 <ProcessOutputs>
 . . . . . . .
  </ProcessOutputs>
</ProcessDescription>
```

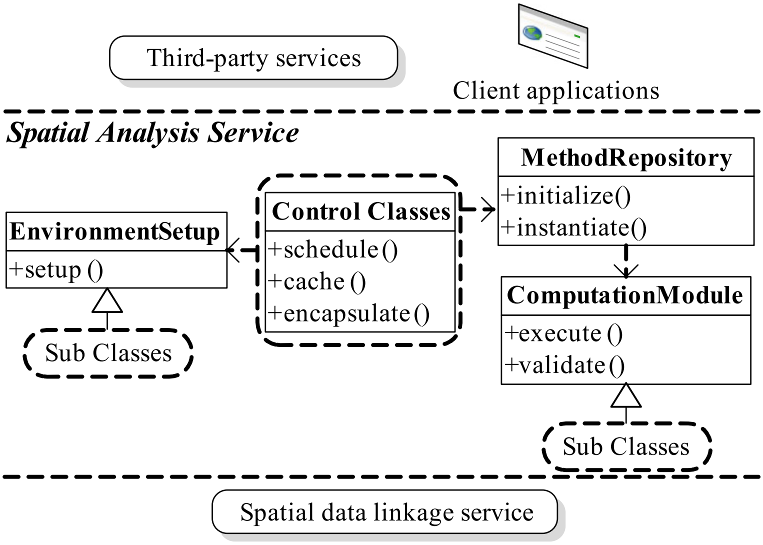


Figure 3 The high-level architecture of the spatial analysis service

The spatial analysis service provides three operations to client applications and services corresponding to a Web Processing Service (WPS): *GetCapabilities*, *DescribeProcess*, and *Execute*. The *DescribeProcess* operation returns a response XML expressing the parameter information including the correlation and validation rules. The output formatted as XML in the *Execute* operation is comprised of computation results as well as the visualization component identifiers for later retrieval.

5.5 The Analysis Workflow

The spatial data linkage and analysis services together provide users with a seamless environment in which to browse and analyze data. Figure 4 shows the workflow of an entire analysis procedure, a typical use of semantic workflows (Gil et al. 2007), applied to a geoportal. The workflow can generally be partitioned into three steps:

Step 1: Users select the data source with interactive selection tools. Users then select a method and fill in the parameters guided by the front-end validation. The request is constructed and sent to the server-side.

Step 2: The server will parse the request; retrieve the required feature set from the spatial data linkage service. After computation, the result is encapsulated into proper inner data structure, and optionally stored in the computation results cache. The service response is then generated and sent to the client after compression.

Step 3: The client-side application parses the server response and extracts the computation results. Users can then interactively explore the analysis results.

6 Implementation

The server-side of the platform is built upon a set of open source software. These include PostgreSQL (Momjian 2001) and PostGIS (PostGIS 2013) for data storage, and GeoServer

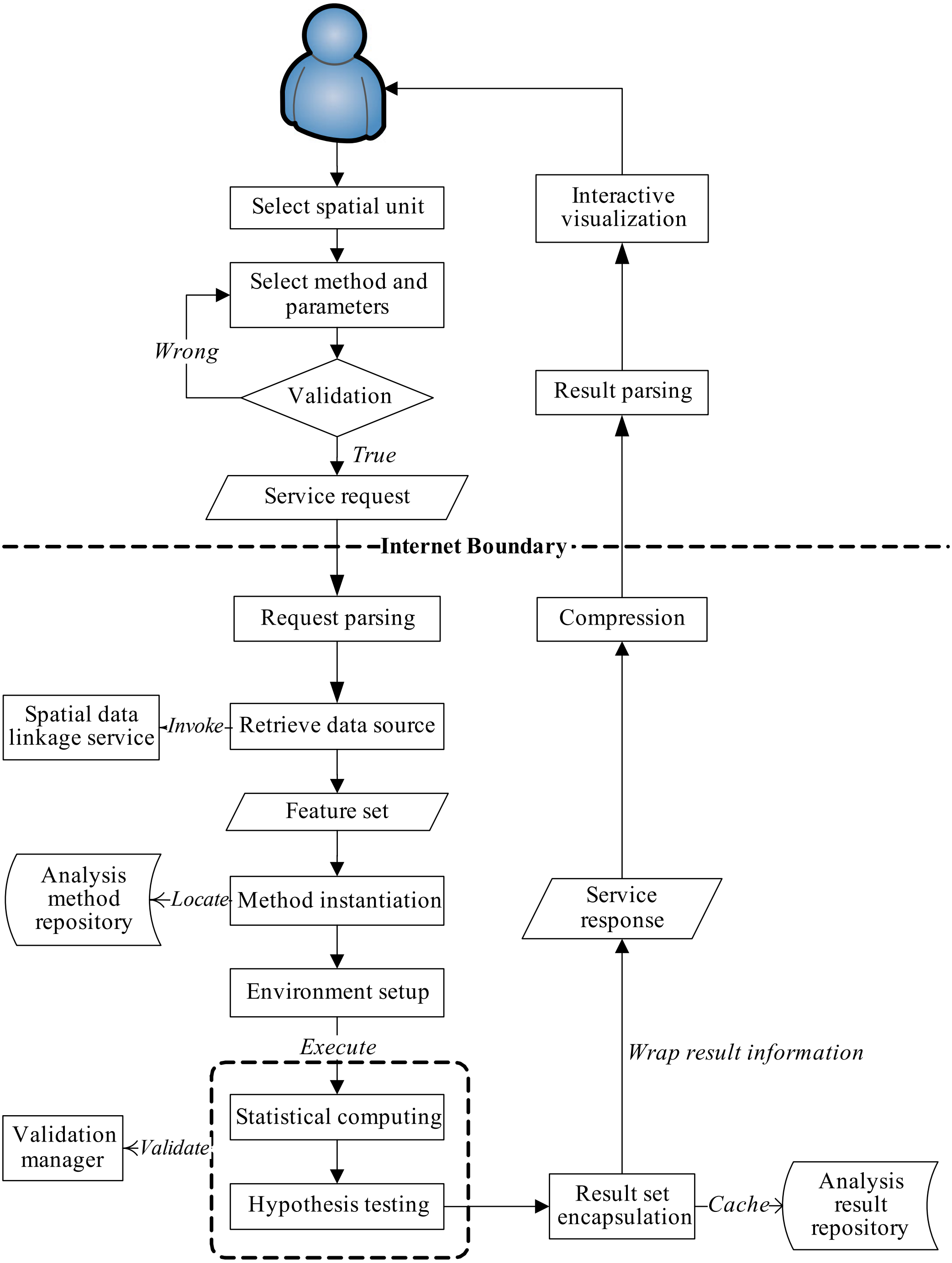


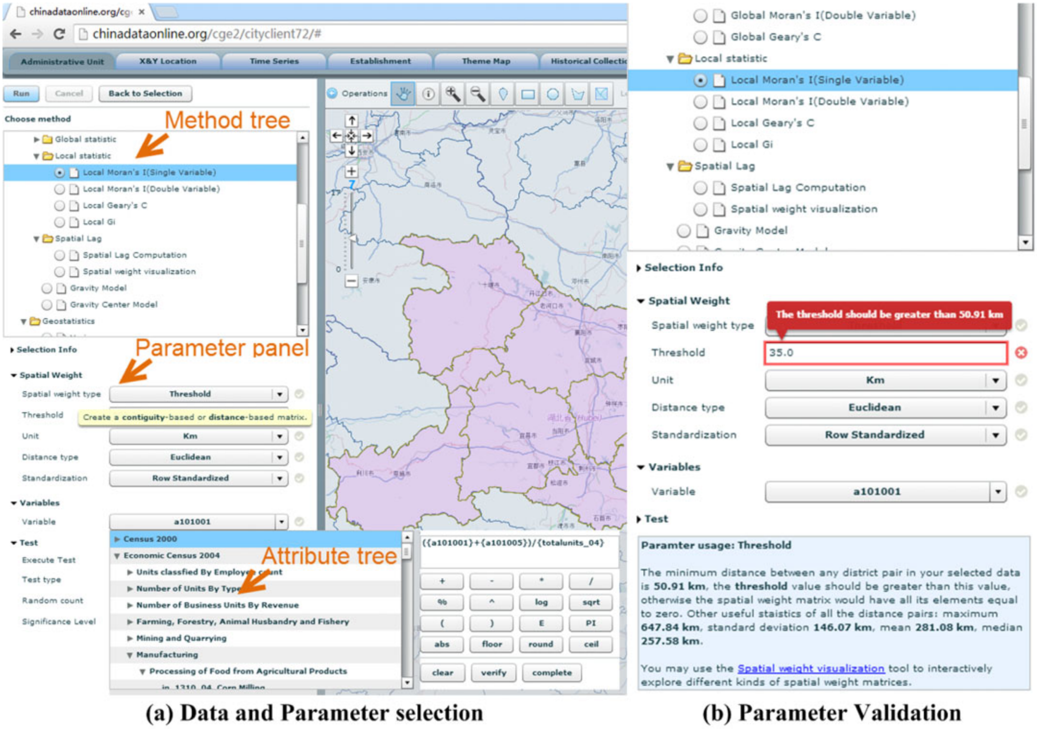
Figure 4 The workflow of an analysis procedure

(Deoliveira 2008) for publishing OGC-compliant services. This technology stack has been commonly used for building geoportals (Moreno-Sanchez et al. 2007). In the geoportal deployment described in this article, the spatial data linkage service integrates 11 linked layers consisting of 51 data sources (WFS data sources or database tables). The spatial analysis service contains a simplified catalogue service for managing different analysis methods. PySAL is invoked through executable Python scripts, while R is invoked through RServe (Urbanek 2003). The web APIs use standard HTTP protocol. The method metadata are parsed on the client side to construct a dynamic form where users can interactively choose parameters and get feedback. Development of the visualization components is supported by a graphics library: StatGL (Data Numerica Institute 2012), which consists of a rich set of interactive plots.

7 Case Study and Discussion

7.1 Exploring Data with Maps, Charts and Reports

Here, we present the ways in which various components in the platform work with the data services to provide functions including spatial analysis, mapping, and reporting. Suppose Miss Li, a graduate student in urban and regional studies, tries to identify the pattern and relationship between household, population, and economic status of Chinese counties. First, Miss Li would like to see the autocorrelation level of each variable using LISA statistics. Figure 5



(a)

Spatial weight type: Threshold

Threshold: Create a contiguity-based or distance-based matrix.

Unit: Km

Distance type: Euclidean

Standardization: Row Standardized

Variable: a101001

({a101001}+{a101005})/{totalunits_04}

(b)

Spatial weight type: The threshold should be greater than 50.91 km

Threshold: 35.0

Unit: Km

Distance type: Euclidean

Standardization: Row Standardized

Variable: a101001

Paramter usage: Threshold

The minimum distance between any district pair in your selected data is **50.91 km**, the **threshold** value should be greater than this value, otherwise the spatial weight matrix would have all its elements equal to zero. Other useful staistics of all the distance pairs: maximum **647.84 km**, standard deviation **146.07 km**, mean **281.08 km**, median **257.58 km**.

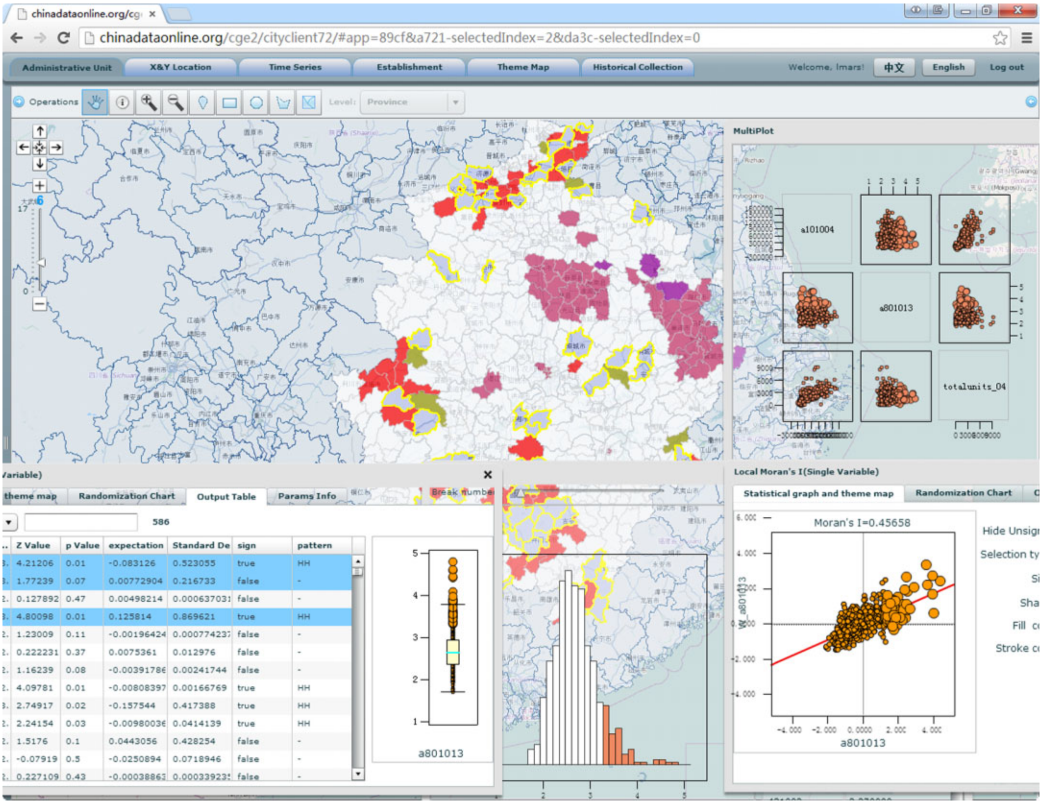
You may use the Spatial weight visualization tool to interactively explore different kinds of spatial weight matrices.

(a) Data and Parameter selection (b) Parameter Validation

Figure 5 Parameter selection and validation of analytic methods

shows an analysis session where she selects all prefecture cities in Hubei Province, and chooses the LISA statistics to explore the spatial autocorrelation. The left function panel shows the method tree and the parameter information. After she selects the method, the lower section of the panel is filled with parameter information with properly pre-defined descriptions, grouping, and allowed values for client-side validation; generated automatically by the corresponding method metadata. The dropdown control prompts Miss Li to choose an existing attribute or manually construct a composite one in the county linked layer. Miss Li was not sure about the parameter selection for the spatial weight matrix, so she tried a threshold-based one, and manually changed the threshold parameter. After she hit the run button, the server-side computation component first validates the input and found the threshold to be too small for the selected data. The service returns a new set of validation rules for the threshold parameter and a message suggesting the parameter usage according to the characteristics of the selected dataset. The minimum value of the updated threshold is the shortest distance among the unit pairs, extracted from the spatial weight matrix of the selected spatial units. The result also includes a link to the spatial weight visualization tool where Miss Li can explore different kinds of spatial weight matrices.

Miss Li continues to explore counties in central China with more exploratory tools (shown in Figure 6). The box plot, histogram, and Moran scatter plot all show the average number of rooms per household, while the multi-plot shows relations between total



586

..	Z Value	p Value	expectation	Standard De	sign	pattern
3.	4.21206	0.01	-0.083126	0.523055	true	HH
3.	1.77239	0.07	0.00772904	0.216733	false	-
2.	0.127892	0.47	0.00498214	0.00063703	false	-
3.	4.80098	0.01	0.125814	0.869621	true	HH
2.	1.23009	0.11	-0.00196424	0.00077423	false	-
2.	0.222231	0.37	0.0075361	0.012976	false	-
2.	1.16239	0.08	-0.0039178	0.00241744	false	-
2.	4.09781	0.01	-0.0080839	0.00166769	true	HH
3.	2.74917	0.02	-0.157544	0.417388	true	HH
2.	2.24154	0.03	-0.0098003	0.0414139	true	HH
2.	1.5176	0.1	0.0443056	0.428254	false	-
2.	-0.07919	0.5	-0.0250894	0.0718946	false	-
2.	0.227109	0.43	-0.0003886	0.0003392	false	-

Figure 6 Exploratory procedures with multiple plots

population, average number of rooms per household, and total economic units. The map, charts, and grid can be linked to help Miss Li better interpret the result.

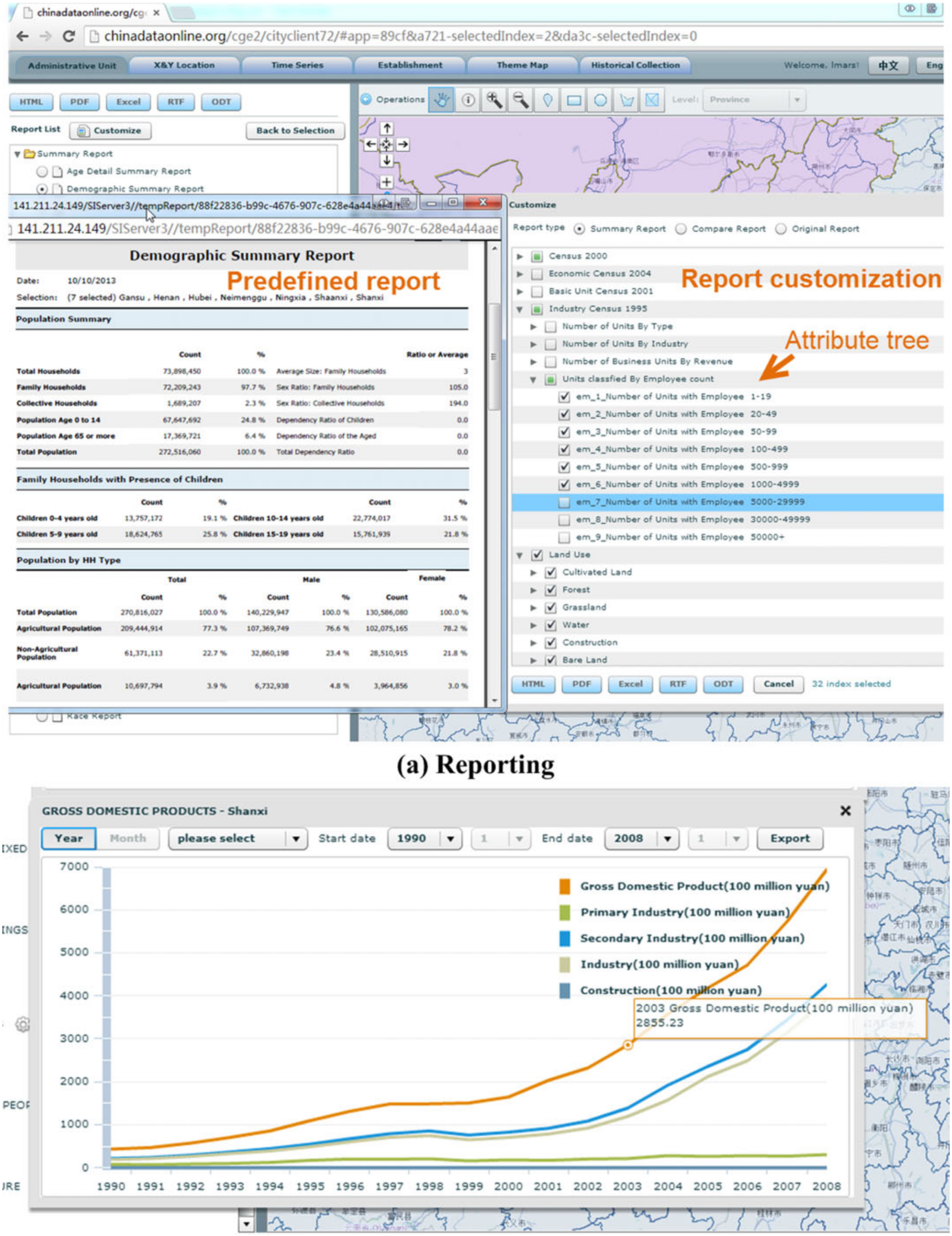
The spatial data linkage service is used across the Geoportal. Figure 7(a) shows a report generated by a predefined report template, as well as the report customization tool that displays the attribute tree for a linked layer. Figure 7(b) show a time-series chart where the client-side application builds the interface according to the meta-information from the virtual attributes, including its temporal scale (yearly in this case), units and number of decimal places.

7.2 Discussion

The spatial data linkage service resembles a database view. Columns from multiple tables are aggregated to several virtual linked layers. The linked layers provide a unified and transparent access to client applications and services. The difference is that they act on the service level, are more flexible with new types of data sources and transformations, and support basic attribute semantics. The data sources currently integrated in the platform mostly have explicit code matching relations. When such codes are missing, a record linkage step that identifies corresponding records within tables (Winkler 2006; Christen 2008) can be plugged in the service by defining composite virtual attributes. This work focuses on a small set of attribute semantics in publishing attributes for urban features. In other applications such as environmental assessment, data quality and uncertainty measures are indispensable attribute semantics. The data type and level of measurement are exposed as the classification of virtual attributes. The service needs to be extended to provide more consistent and coherent meta-attributes to represent classification of attributes. Currently, administrators are required to fully involve in the process of defining linked layers and their virtual attributes. Making this process more automatic and intelligent is important for the service to integrate more diverse datasets in a timely and scalable manner. Our spatial data linkage service is a lightweight solution for Web-based mapping applications, and causes little disruption to existing data services. With carefully defined geospatial ontologies and properly implemented query APIs, the spatial data linkage service can serve as a RDB-to-RDF mapping tier (Hert et al. 2011) to represent geospatial data more comprehensively. To familiarize users quickly with the increasing number of analysis methods available, our analysis service is designed to present parameters in an informative and organized way, minimize the chances of erroneous input, and provide meaningful feedback when errors do happen. The enriched parameter expression incorporated in the analytic services reduces the ambiguity by restriction and validation of input values, trying to strike a balance between interoperability, comprehensiveness, and interactivity. Standardization might be an issue when extending this approach to other application areas. Providing a full and standard classification of spatial analytical methods requires extensive negotiations and collaboration. These collaborative efforts will be of great value, since many spatial analytic methods are commonly used in various disciplines that have a spatial dimension.

8 Conclusions

Recent years have seen a growing interest in building user-centric and visual-rich workbenches that integrate heterogeneous data and tools, while at the same time maintaining the interoperability and scalability through adherence to standard specifications and data formats. This article describes a flexible architecture for spatial data linkage and analysis services that



▼ Summary Report
○ Age Detail Summary Report
● Demographic Summary Report

141.211.24.149/SIServer3//tempReport/88f22836-b99c-4676-907c-628e4a44aae4/t

141.211.24.149/SIServer3//tempReport/88f22836-b99c-4676-907c-628e4a44aae

Demographic Summary Report

Date: 10/10/2013

Selection: (7 selected) Gansu , Henan , Hubei , Neimenggu , Ningxia , Shaanxi , Shanxi

Population Summary

	Count	%		Ratio or Average
Total Households	73,898,450	100.0 %	Average Size: Family Households	3
Family Households	72,209,243	97.7 %	Sex Ratio: Family Households	105.0
Collective Households	1,689,207	2.3 %	Sex Ratio: Collective Households	194.0
Population Age 0 to 14	67,647,692	24.8 %	Dependency Ratio of Children	0.0
Population Age 65 or more	17,369,721	6.4 %	Dependency Ratio of the Aged	0.0
Total Population	272,516,060	100.0 %	Total Dependency Ratio	0.0

Family Households with Presence of Children

	Count	%		Count	%
Children 0-4 years old	13,757,172	19.1 %	**Children 10-14 years old**	22,774,017	31.5 %
Children 5-9 years old	18,624,765	25.8 %	**Children 15-19 years old**	15,761,939	21.8 %

Population by HH Type

	Total		Male		Female	
	Count	%	Count	%	Count	%
Total Population	270,816,027	100.0 %	140,229,947	100.0 %	130,586,080	100.0 %
Agricultural Population	209,444,914	77.3 %	107,369,749	76.6 %	102,075,165	78.2 %
Non-Agricultural Population	61,371,113	22.7 %	32,860,198	23.4 %	28,510,915	21.8 %
Agricultural Population	10,697,794	3.9 %	6,732,938	4.8 %	3,964,856	3.0 %

Customize

Report type ● Summary Report ○ Compare Report ○ Original Report

- ▶ Census 2000
- ▶ □ Economic Census 2004
- ▶ □ Basic Unit Census 2001
- ▼ Industry Census 1995
  - ▶ □ Number of Units By Type
  - ▶ □ Number of Units By Industry
  - ▶ □ Number of Business Units By Revenue
  - ▼ Units classfied By Employee count
    - ☑ em_1_Number of Units with Employee 1-19
    - ☑ em_2_Number of Units with Employee 20-49
    - ☑ em_3_Number of Units with Employee 50-99
    - ☑ em_4_Number of Units with Employee 100-499
    - ☑ em_5_Number of Units with Employee 500-999
    - ☑ em_6_Number of Units with Employee 1000-4999
    - □ em_7_Number of Units with Employee 5000-29999
    - □ em_8_Number of Units with Employee 30000-49999
    - □ em_9_Number of Units with Employee 50000+
- ▼ ☑ Land Use
  - ▶ ☑ Cultivated Land
  - ▶ ☑ Forest
  - ▶ ☑ Grassland
  - ▶ ☑ Water
  - ▶ ☑ Construction
  - ▶ ☑ Bare Land

(a) Reporting

GROSS DOMESTIC PRODUCTS - Shanxi

Year
Month
please select
Start date
1990
1
End date
2008
1
Export
7000
6000
5000
4000
3000
2000
1000
0
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Gross Domestic Product(100 million yuan)
Primary Industry(100 million yuan)
Secondary Industry(100 million yuan)
Industry(100 million yuan)
Construction(100 million yuan)
2003 Gross Domestic Product(100 million yuan)
2855.23
IXED
INGS
PEOF
JRE

(b) Time-series data display

Figure 7 Reporting and time-series data display

are integrated in the online platform for China urban research. The spatial data linkage service acts as a middleware for multi-source spatiotemporal data, and provides client applications and services with unified and transparent access APIs. The spatial analysis service explicitly incorporates method and parameter semantics in the metadata and the computation components to assist users in operating and understanding the available analytic methods. Future work will investigate automatic construction of linked layers from contextual information that coexist with the published data services; fit these two services into the semantic web; integrate a task-oriented architecture using a standard workflow engine to chain the atomic Web Processing Service (WPS) elements; and leverage computing facilities in CyberGIS (Wang 2010) to support high performance data analysis. We also plan to integrate volunteered geographic information (Goodchild 2007) such as individual trajectories to adapt the platform to broader applications.

References

Amirian P, Alesheikh A A, and Bassiri A 2010 Standards-based, interoperable services for accessing urban services data for the city of Tehran. *Computers, Environment and Urban Systems* 34: 309–21

Andrienko G, Andrienko N, Demsar U, Dransch D, Dykes J, Fabrikant S I, Jern M, Kraak M-J, Schumann H, and Tominski C 2010 Space, time and visual analytics. *International Journal of Geographical Information Science* 24: 1577–600

Andrienko N and Andrienko G 2006 *Exploratory Analysis of Spatial and Temporal Data*. Berlin, Springer

Anselin L 1995 Local Indicators of Spatial Association – LISA. *Geographical Analysis* 27: 93–115

Anselin L, Kim Y W, and Syabri I 2004 Web-based analytical tools for the exploration of spatial data. *Journal of Geographical Systems* 6: 197–218

Anselin L and Rey S J 2012 Spatial econometrics in an age of CyberGIScience. *International Journal of Geographical Information Science* 26: 2211–26

Auer T, MacEachren A M, McCabe C, Pezanowski S, and Stryker M 2011 HerbariaViz: A web-based client–server interface for mapping and exploring flora observation data. *Ecological Informatics* 6: 93–110

Batcheller J K and Reitsma F 2010 Implementing feature level semantics for spatial data discovery: Supporting the reuse of legacy data using open source components. *Computers, Environment and Urban Systems* 34: 333–44

Berkley C, Bowers S, Jones M B, Ludäscher B, Schildhauer M, and Tao J 2005 Incorporating semantics inscientific workflow authoring. In *Proceedings of the Seventeenth International Conference on Scientific and Statistical Database Management*, Berkeley, California: 75–8

Christen P 2008 Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada: 151–59

Data Numerica Institute 2012 StatGL Graphical Library: A foundation for online graphical data analysis. WWW document, http://datanumerica.com/StatGL.html

Deelman E, Gannon D, Shields M, and Taylor I 2009 Workflows and e-Science: An overview of workflow system features and capabilities. *Future Generation Computer Systems* 25: 528–40

Deoliveira J 2008 GeoServer: Uniting the GeoWeb and spatial data infrastructures. In *Proceedings of the Tenth International Conference for Spatial Data Infrastructure*, St. Augustine, Trinidad

Fonseca F T, Egenhofer M J, Agouris P, and Câmara G 2002 Using ontologies for integrated geographic information systems. *Transactions in GIS* 6: 231–57

Frew J and Bose R 2001 Earth System Science Workbench: A data management infrastructure for earth science products. In *Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM 2001)*, Fairfax, Virginia: 180–9

Frew J, Freeston M, Freitas N, Hill L, Janée G, Lovette K, Nideffer R, Smith T, and Zheng Q 2000 The Alexandria digital library architecture. *International Journal on Digital Libraries* 2: 259–68

Fry R, Berry R, Higgs G, Orford S, and Jones S 2012 The WISERD Geoportal: A tool for the discovery, analysis and visualization of socio-economic (meta-) data for Wales. *Transactions in GIS* 16: 105–24

Ghaemi P, Swift J, Sister C, Wilson J P, and Wolch J 2009 Design and implementation of a web-based platform to support interactive environmental planning. *Computers, Environment and Urban Systems* 33: 482–91

Gil Y, Deelman E, Ellisman M, Fahringer T, Fox G, Gannon D, Goble C, Livny M, Moreau L, and Myers J 2007 Examining the challenges of scientific workflows. *Computer* 40(12): 24–32

Gil Y, Ratnakar V, and Fritz C 2010 Assisting scientists with complex data analysis tasks through semantic workflows. In *Proceedings of the AAAI Fall Symposium on Proactive Assistant Agents*, Arlington, Virginia

Gil Y, Ratnakar V, Kim J, González-Calero P, Groth P, Moody J, and Deelman E 2011 Wings: Intelligent workflow-based design of computational experiments. *Intelligent Systems* 26(1): 62–72

Goodchild M, Buttenfield B, Adler P, Krygiel A, Onsrud H, and Kahn R 1999 *Distributed Geolibraries*. Washington, DC, National Academies Press

Goodchild M F 2007 Citizens as sensors: The world of volunteered geography. *GeoJournal* 69: 211–21

Hert M, Reif G and Gall H C 2011 A comparison of RDB-to-RDF mapping languages. In *Proceedings of the Seventh ACM International Conference on Semantic Systems*, Graz, Austria: 25–32

Hey T and Trefethen A E 2005 Cyberinfrastructure for e-Science. *Science* 308(5723): 817–21

Ho Q, Lundblad P, Åström T, and Jern M 2012 A web-enabled visualization toolkit for geovisual analytics. *Information Visualization* 11(1): 22–42

Kumar V, Kurc T, Ratnakar V, Kim J, Mehta G, Vahi K, Nelson Y, Sadayappan P, Deelman E, Gil Y, Hall M, and Saltz J 2010 Parameterized specification, configuration and execution of data-intensive scientific workflows. *Cluster Computing* 13: 315–33

Li S, Xiong C, and Ou Z 2011 A Web GIS for sea ice information and an ice service archive. *Transactions in GIS* 15: 189–211

MacEachren A M, Crawford S, Akella M, and Lengerich G 2008 Design and implementation of a model, Web-based, GIS-enabled cancer atlas. *Cartographic Journal* 45: 246–60

Momjian B 2001 *PostgreSQL: Introduction and Concepts*. New York, Addison-Wesley

Moreno-Sanchez R, Anderson G, Cruz J, and Hayden M 2007 The potential for the use of Open Source Software and Open Specifications in creating Web-based cross-border health spatial information systems. *International Journal of Geographical Information Science* 21: 1135–63

OGC 2012a *OGC GeoSPARQL-A Geographic Query Language for RDF Data*. Wayland, MA, OGC Candidate Implementation Standard

OGC 2012b *OGC® Geography Markup Language (GML): Extended Schemas and Encoding Rules*. Wayland, MA, OGC Implementation Standard

Pettit C, Widjaja I, Russo P, Sinnott R, Stimson R, and Tomko M 2012 Visualisation support for exploring urban space and place. *ISPRS Annals of Photogrammetry, Remote Sensing, and Spatial Information Science* I-2: 153–58

PostGIS 2013 About PostGIS. WWW document, http://postgis.net/

R Development Core Team 2011 *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing

Rey S J 2009 Show me the code: Spatial analysis and open source. *Journal of Geographical Systems* 11: 191–207

Rey S J and Anselin L 2010 PySAL: A Python library of spatial analytical methods. In Fischer M M and Getis A (eds) *Handbook of Applied Spatial Analysis,* Berlin, Springer: 175–93

Schut P and Whiteside A 2007 *OpenGIS Web Processing Service*. Wayland, MA, OGC Project Document

She B, Zhu X, and Xiao W 2012 Building an integrated web-based environment for exploratory spatiotemporal data analysis. *ISPRS Annals of Photogrammetry, Remote Sensing, and Spatial Information Science* I-4: 169–74

Sinnott R O, Galang G, Tomko M, and Stimson R 2011 Towards an e-infrastructure for urban research across Australia. In *Proceedings of the IEEE e-Science Conference*, Stockholm, Sweden: 295–302

Stadler C, Lehmann J, Höffner K, and Auer S 2012 LinkedGeoData: A core for a web of spatial open data. *Semantic Web* 3: 333–54

Stevens S S 1946 On the theory of scales of measurement. *Science* 103(2684): 677–80

Takatsuka M and Gahegan M 2002 GeoVISTA Studio: A codeless visual programming environment for geoscientific data analysis and visualization. *Computers and Geosciences* 28: 1131–44

Tomko M, Greenwood P, Sarwar M, Morandini L, Stimson R, Bayliss C, Galang G, Nino-Ruiz M, Voorsluys W, Widjaja I, Koetsier G, Mannix D, Pettit C, and Sinnott R 2012 The design of a flexible web-based analytical platform for urban research. In *Proceedings of the Twentieth International Conference on Advances in Geographic Information Systems*, Redondo Beach, California: 369–75

Urbanek S 2003 Rserve: A fast way to provide R functionality to applications. In *Proceedings of the Third International Workshop on Distributed Statistical Ccomputing (DSC 2003)*, Vienna, Austria

Valentine D, Taylor P and Zaslavsky I 2012 WaterML, an information standard for the exchange of in-situ hydrological observations. In *Proceedings of the European Geosciences Union General Assembly*, Vienna, Austria: 13275

Vivid Solutions 2013 JTS Topology Suite. WWW document, http://www.vividsolutions.com/jts/JTSHome.htm

Wang S 2010 A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers* 100: 535–57

Wang S, Wilkins-Diehr N R, and Nyerges T L 2012 CyberGIS: Toward synergistic advancement of cyberinfrastructure and GIScience: A workshop summary. *Journal of Spatial Information Science* 4: 125–48

Winkler W E 2006 *Overview of Record Linkage and Current Research Directions*. Washington, DC, US Bureau of the Census Technical Report Statistical Research Report Series RRS2006/02

Yang C, Raskin R, Goodchild M, and Gahegan M 2010 Geospatial Cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems* 34: 264–77

Yue P, Gong J, Di L, Yuan J, Sun L, Sun Z and Wang Q 2010 GeoPW: Laying blocks for the geospatial processing web. *Transactions in GIS* 14: 755–72

Yue P, Wei Y, Di L, He L, Gong J, and Zhang L 2011 Sharing geospatial provenance in a service-oriented environment. *Computers, Environment and Urban Systems* 35: 333–43

Zhang C, Zhao T, Li W, and Osleeb J P 2010 Towards logic-based geospatial feature discovery and integration using web feature service and geospatial semantic web. *International Journal of Geographical Information Science* 24: 903–23