

## MAKING SPATIAL STATISTICS SERVICE ACCESSIBLE ON CLOUD PLATFORM

Xiaoyan Mu <sup>1,\*</sup>, Jingyang Wu <sup>1</sup>, Te Li <sup>1</sup>, Ying Zhong <sup>1</sup>, Xiang Gao <sup>1</sup>

<sup>1</sup> School of Resource and Environmental Science of Wuhan University, 129 Luoyu Road, Wuhan 430079, China

\*Corresponding author, E-mail: [icgs@whu.edu.cn](mailto:icgs@whu.edu.cn)

**KEY WORDS:** Web service, Cloud platform, Microsoft Windows Azure, Eigenvector-based spatial filtering, Regression analysis, DLL

### ABSTRACT:

Web service can bring together applications running on diverse platforms, users can access and share various data, information and models more effectively and conveniently from certain web service platform. Cloud computing emerges as a paradigm of Internet computing in which dynamical, scalable and often virtualized resources are provided as services. With the rampant growth of massive data and restriction of net, traditional web services platforms have some prominent problems existing in development such as calculation efficiency, maintenance cost and data security. In this paper, we offer a spatial statistics service based on Microsoft cloud. An experiment was carried out to evaluate the availability and efficiency of this service. The results show that this spatial statistics service is accessible for the public conveniently with high processing efficiency.

### 1. INTRODUCTION

Web service can bring together applications running on diverse platforms, enable database information exchange and allow applications originally meant for internal use be made available through the Internet (Christopher Ferris & Joel Farrell 2003). Cloud computing emerges as a paradigm of Internet computing in which dynamical, scalable and often virtualized resources are provided as services. With the rampant growth of massive data and restriction of net, traditional web services platforms have some prominent problems existing in development such as calculation efficiency, maintenance cost and data security. To solve these problems, individual applications and enterprise storage are being deployed on clouds. By deploying service in the cloud, such infrastructure as storage and networks, which needs to be supplied by the developer in a traditional host-client pattern, is now available through the Internet from anywhere.

In this paper, we propose some spatial statistics services such as matrix operations service, statistical calculation service and regression analysis service. Integrating these services, we report an effort to develop a service system providing eigenvector-based spatial filtering (ESF) method to perform regression analysis with spatial data. This ESF service system makes users access to the powerful regression method by provisioning a series of concerned services with the support of cloud computing. With these services, users can efficiently and conveniently perform regression analysis with their own spatial data. Microsoft Windows Azure is chosen as the platform to develop and deploy these services and the ESF service system. Microsoft Windows Azure actually consists of a variety of different services on top of a common platform and it supports

applications built on the .NET Framework and other popular programming languages supported in Windows systems (Zhang et al., 2010). Azure is selected as the platform for the ESF service due to our familiarity with the .NET framework and the development tools. Azure provides the full range services for developing the ESF system that takes advantage of parallel computing and easy access of the Web.

The remainder of this paper is arranged as follows: the second section analyzes the algorithm of ESF. Section three designs the service system and Web page layout. Subsequently, an experiment is presented to demonstrate the availability and accessibility of these services and the accuracy of the eigenvector-based spatial filtering regression model.

### 2. ANALYSIS OF ESF ALGORITHM

Ordinary Least Square (OLS) is a commonly used method in regression analysis. However, when dealing with spatial data, the existence of spatial autocorrelation often leads to violations of the basic assumptions for OLS (Tobler, 1970; Hubert et al., 1981; Getis, 1990). The eigenvector-based spatial filtering (ESF) service provides an effective spatial regression model to handle with spatial autocorrelations. The effect of the variables in the regression model is divided into two parts: spatial influence and non-spatial influence. Extracting the spatial influence part and filtering it, then we can use the basic method to make regression analysis.

---

About the first author: Mu Xiaoyan, E-mail: [icgs@whu.edu.cn](mailto:icgs@whu.edu.cn)

## 2.1 ESF algorithm process

- (1) Generate the spatial connectivity matrix.
- (2) Calculate the eigenvalue and eigenvector of the connectivity matrix.
- (3) Select eigenvectors as spatial filters.
- (4) Perform Ordinary Least Square regression analysis with the original explanatory variables and the selected eigenvectors.
- (5) Assess the quality of the model.

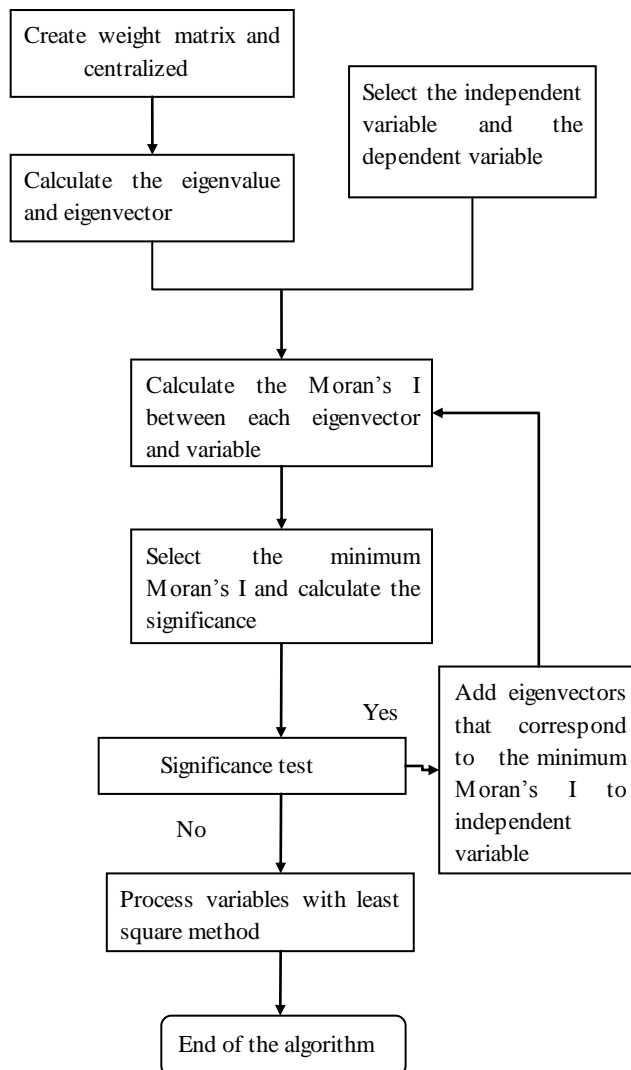


Figure 1. Algorithm flow diagram

## 2.2 The concrete implementation of the ESF algorithm

Component Object Model (COM) from Microsoft is one of the mature component technologies and widely employed in its operating system and application software. Component technology is a powerful and hot scheme to actualize software reuse, shorten programming time, decrease maintenance costs and realize dynamic program upgrade.

Dynamic Link Library (DLL) provides a convenient way to share data and code. It can save disk space significantly for different applications can share the same DLL function. Various

of programming tools can be implemented to write DLL files such as vc++, c++ Builder and Delphi.

Eigenvector-based spatial filtering algorithm is written with COM technology in Microsoft Visual C++ 6.0 environment. Then the corresponding DLLs are generated to be accessed dynamically at run time and invoked by the share service.

### 2.2.1 The overall framework and functional module

The ESF service system will be divided into four major modules: fundamental function module, matrix calculation module, statistical computing module and regression calculation module. There are four classes according to the module division. The fundamental functional class is not required as the interface exposed to the users, so it is a fundamental class while the other three classes are the interface classes. The next section will give detailed descriptions of these classes.

### 2.2.2 The design of the class

(1) CComBase is a common class to achieve fundamental functions. It contains some functions for the subsequent calculation.

(2) CmatrixBase is a class for matrix operations, and ultimately exposed to the user with the form of interface. Users can invoke external without understanding details of its implementation. The interfaces and functions are shown in table 1.

Interface Name	Interface function
Create_Weight	Create weight matrix
Weight_Centered	Center weight matrix
Variable_Centered	Center variables
Transpose_Matrix	Calculate transpose matrix
Save_Matrix	Storage matrix
Normal_Matrix	Calculate normal matrix
Merge_Matrix	Merge matrixes
Inverse_Matrix	Calculate inverse matrix
Dsyevr_Eigen	Calculate the eigenvalue and eigenvector

Table 1. Introduction for the CmatrixBase interface and function  
(3) CStatistic is an interface class for statistical computing function. It implements intermediary step of the ESF algorithm and can be invoked by users. The interfaces and functions are shown in table 2.

Interface Name	Interface function
Hat_Matrix	Calculate hat matrix
Cal_Residual	Calculate Residual
Cal_MoranI	Calculate Moran's I
OLS_Regression	Ordinary Least Square Regression

Table 2. Introduction for the CStatistic interface and function  
(4) Ceigenfunction is an interface class for regression computing. It is important step of the ESF algorithm and can be invoked by users. The interfaces and functions are shown in table 3.

Interface Name	Interface function
ReadFieldsName	Get attribute names
Simple_ratio	Choose eigenvector with Simple_ratio method
Step_wiseV	Choose eigenvector with Step_wise method
Regression	Regression analysis

Table 3. Introduction for the Ceigenfunction interface and function

### 2.2.3 The generation of DLL

Build function framework for COM component according to the system structure designed in the previous section in visual c++ 6.0. So we can have more clearly organized to write code. After the completion of code editor, click the DLL generation button in the Build TAB, then complete the generation of a DLL. Users can invoke the interface functions in other programs after registering the DLL on their computers.

## 3. RESEARCH OF ESF SERVICE SYSTEM

### 3.1 The system design of ESF service

Making DLL component function in to web service, and then deploy the web service on cloud platform for users to access. Edit the web service interface for invoke to achieve the function of regression analysis. Specific implementation process is shown in figure 2.

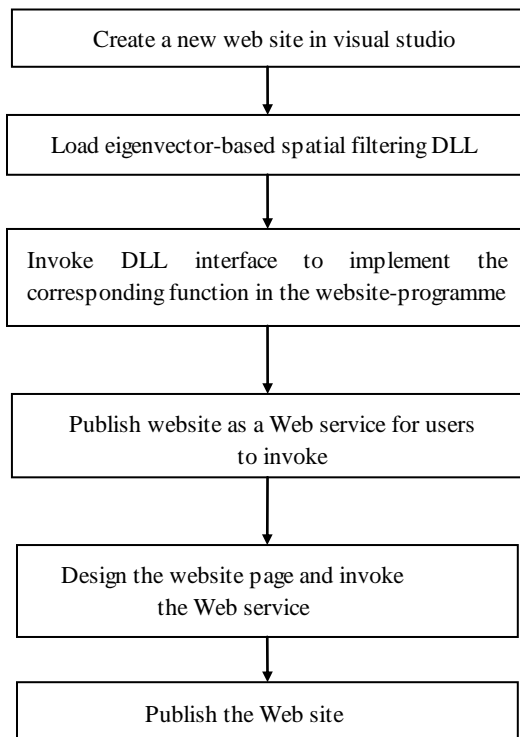


Figure 2. Web service process

ESF service system is a more complete service system, including user management, document management and data processing and other functions. Service architecture is shown in figure 3.

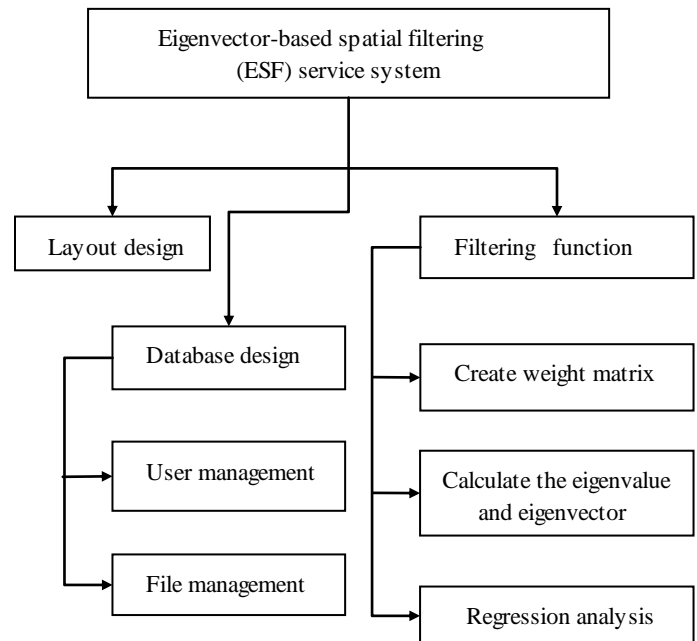


Figure 3. Service architecture

### 3.2 The service design of ESF service

Use ASP.NET to compile ESF algorithm DLL into the Web Service interface in the Microsoft Visual Studio 2010 environment. And then invoke the web service in website to allow the users to complete regression analysis. The main interfaces and functions are shown in table 4.

Interface Name	Interface function
Cal_MoranI	Calculate Moran's I
CalculateEigen	Calculate the eigenvalue and eigenvector
Hat_Matrix	Calculate hat matrix
OLS_Regression	Ordinary Least Square Regression
RegressionModel	Regression analysis
createWFilebyShp	Add attributes to DBF database
getDBFfieldName	Get DBF attribute name
simpleSelect	Simple selection method to choose eigenvector

Table 4. Introduction for web service interface and function

### 3.3 The website design of the service

We create two websites (web service website and ESF service website which integrates related web services) considering that users may need one or more services.

The name of the interface corresponds to its function in the web service website. Each web service has some brief instructions on the interface so that users can have a clearly understanding of its use.



Figure 4. Web service site



Figure 5. The portal site of ESF service system

## 4. EXPERIMENT

### 4.1 Experiment data

Data content: Economic development status and other basic attributes about thirty provinces of Chinese.

The Data used in the experiment contains the dependent variable GDP and independent variable AREA and Pop.

The detailed data is shown in figure 6.

F1	B	C	D	E	F	G	H	I
	PERIMETER	AREA	ADCODE93	ADCODE99	NAME	GDP	Pop	SFILTER
1	68.489	54.447	230000	230000	黑龙江省	8257.24000	3824	685.2867000
2	129.933	129.113	150000	150000	内蒙古自治区	8967.52000	2405	619.2514000
3	84.905	175.591	650000	650000	新疆维吾尔自治区	4005.41000	2095	-478.6630000
4	41.186	21.315	220000	220000	吉林省	7072.25000	2730	918.8021000
5	38.379	15.603	210000	210000	辽宁省	14696.20000	4298	870.5908000
6	76.781	41.508	620000	620000	甘肃省	3373.78000	2619	-616.7130000
7	44.874	19.504	130000	130000	河北省	17068.00000	6943	505.5459000
8	8.498	1.733	110000	110000	北京市	11469.30000	1633	484.3685000
9	24.784	15.961	140000	140000	山西省	7050.38000	3393	-49.9480000
10	7.263	1.214	120000	120000	天津市	7068.56000	1115	484.3685000
11	38.571	20.394	610000	610000	陕西省	7752.20000	3748	-788.9350000
12	18.520	5.272	640000	640000	宁夏回族自治区	1198.15000	610	-166.7250000
13	59.562	71.363	630000	630000	青海省	1012.69000	552	-631.0060000
14	40.900	15.388	370000	370000	山东省	33621.30000	9367	-557.2750000
15	76.629	114.331	540000	540000	西藏自治区	434.34000	284	-389.6980000
16	30.911	16.135	410000	410000	河南省	19724.70000	9360	-803.6700000
17	27.893	9.736	320000	320000	江苏省	33478.80000	7625	-593.4000000
18	29.492	13.365	340000	340000	安徽省	10191.50000	6118	-960.3650000
19	58.636	45.534	510000	510000	四川省	14050.80000	8127	-567.9730000
20	34.321	17.563	420000	420000	湖北省	12866.00000	5699	-673.1230000
21	26.012	7.716	510000	500000	重庆市	5693.58000	2816	-224.9980000
22	26.825	9.277	330000	330000	浙江省	22717.00000	5060	-336.8220000
23	32.801	19.390	430000	430000	湖南省	12939.80000	6355	597.3233000
24	26.668	15.277	360000	360000	江西省	6854.12000	4368	12.6938000

Figure 6. Experiment data

### 4.2 The experiment process

#### 4.2.1 Upload personal data

To use these services, users may need to register a personal account at first. After the success of the registration, you can upload spatial data to the server side for spatial statistics service. In the individual data centre, practical directory services including file selection, file deletion, file downloading and file sharing are provided for the convenience of data management.

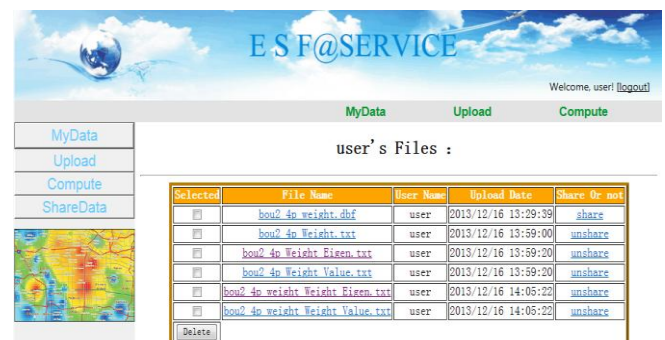


Figure 7. Personal data center

#### 4.2.2 Generate spatial connectivity matrix by shp

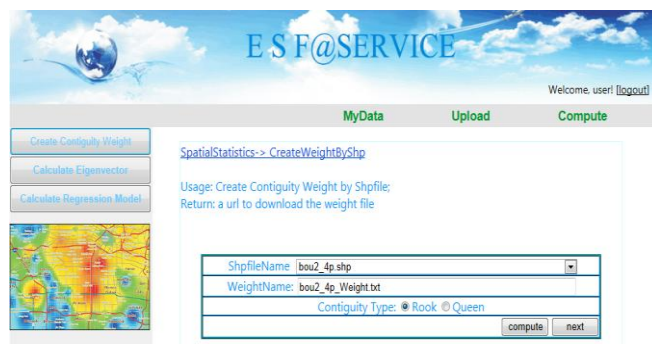


Figure 8. The result of connectivity matrix

#### 4.2.3 Calculating the eigenvalue and eigenvector



Figure 9. The result of eigenvector and eigenvalue

#### 4.2.4 Regression analysis

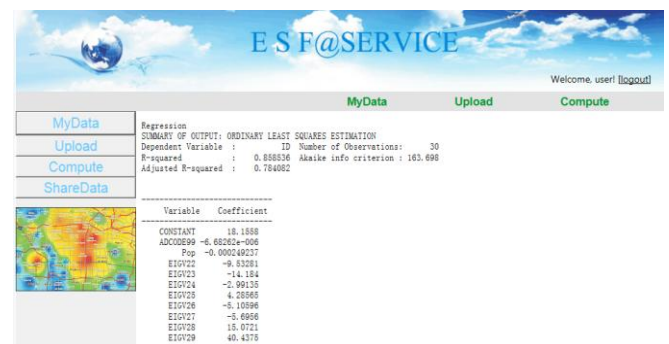


Figure 10. The result of regression analysis

From the friendly user and result interfaces showed above, it can be seen that ESF service is accessible for the public conveniently with high processing efficiency. In addition, deployed on Microsoft cloud, these spatial statistics services can be remote accessed by entering website in to browser address bar wherever and whenever.

#### 4.3 Comparison of different regression results

Area, population and GDP of each province are selected as experiment data. Perform regression analysis with this data respectively in Geoda software and ESF system. The different regression results are shown in table 5.

	Y	X	Numbers of Select eigen	R-squared	Adjusted R-squared	Akaike info criterion
Geo da- Clas sic	G D P	AR EA Pop		0.640175	0.613521	609.068
ESF - Sim ple	G D P	AR EA Pop	8	0.838517	0.753526	580.476

Table 5. Regression results in different software

Illustration: Geoda - Classic: regression result with Geoda software; ESF – Simple (simple ratio to choose eigenvector): regression result with ESF system.

In all the regressions, four parameters are calculated to assess the quality of those models. For the convenience of comparison, the significant tests for regression equation and regression parameters are not listed here. Among those parameters, both R-squared and Adjusted R-squared are used to test the goodness of fit of the regression model. A low number means worse fitting degree while higher numbers means good fit degree. Akaike info criterion is used to test the accuracy of the regression model. A low number means good accuracy while higher numbers means worse.



Result: As displayed in the table above, ESF service system has better fitting degree than Geoda software and improves the accuracy of the regression model.

## 5. CONCLUSION

This paper proposed an ESF service system and some statistics services based on a cloud platform, Windows Azure. By deploying the system on the cloud, the availability and accessibility of these services have been significantly improved. With this system, users can conveniently get access to the spatial filtering method and use those services to analyze spatial data efficiently. Eigen function based spatial filtering method can eliminate the spatial autocorrelation to some extent and improve the accuracy of the regression model.

## 6. ACKNOWLEDGEMENT

The research is supported by grants from National Nature Science Foundation of China (project No. 41171347).

## 7. REFERENCES

- [1] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1), 7-18.
- [2] Kun, Y., WANG, X., & ZHOU, R. (2004). Underlying Techniques for Web Services: A Survey. *Journal of Software*, Vol.15, No.3.
- [3] Huhns, M.N., Singh, M.P. (2005). Service-oriented computing: key concepts and principles. *IEEE Internet Computing*, vol. 09, pp. 75 - 81.
- [4] Ferris, C., Farrell, J. (2003). What are Web Services? *Communications of the ACM*, vol. 46, No.6 pp. 31.
- [5] Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), 234-240.
- [6] Hubert, L. J., Golledge, R. G., & Costanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, 13(3), 224-233.
- [7] Getis, A. (1990). Screening for spatial dependence in regression analysis. *Papers in Regional Science*, 69(1), 69-81.
- Griffith, D. A. (2000). A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems*, 2(2), 141-156.
- [8] Griffith, D. A., & Peres-Neto, P. (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, 87(10), 2603-2613.
- [9] Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., ... & Stoica, I. (2009). Above the clouds: A Berkeley view of cloud computing. Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS, 28.
- [10] Johnston, S. J., O'Brien, N. S., Lewis, H. G., Hart, E. E., White, A., & Cox, S. J. (2013). Clouds in space: Scientific computing using windows azure. *Journal of Cloud Computing*, 2(1), 1-10.
- [11] Molto, G., Calatrava, A., & Hernandez, V. (2013). A Service-Oriented Architecture for Scientific Computing on Cloud Infrastructures. In M. Dayde, O. Marques, & K. Nakajima (Eds.), *VECPAR 2012* (pp. 163-176). Berlin: Springer.
- [12] Shiraz, M., Abolfazli, S., Sanaei, Z., & Gani, A. (2013). A study on virtual machine deployment for application outsourcing in mobile cloud computing. *The Journal of Supercomputing*, 63(3), 946-964.
- [13] Häfner, C. N., & Karagiannis, G. (2011). Cloud computing services: taxonomy and comparison. *Journal of Internet Services and Applications*, 2(2), 81-94.