# A Web-based Framework for Visualizing Industrial Spatiotemporal Distribution using Standard Deviational Ellipse and Shifting Routes of Gravity Centers

Yunting Song [a], Zhipeng Gui [a,b,c] *, Huayi Wu [b,c], Yangjiaxin Wei[d]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China – (dreamcloud0704, zhipeng.gui) @whu.edu.cn
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China – (zhipeng.gui, wuhuayi) @whu.edu.cn
[c] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China – (zhipeng.gui, wuhuayi) @whu.edu.cn
[d] Department of Geography, University of Georgia, Athens, GA, USA – yw27596@uga.edu

**Commission V, WG V/4**

**KEY WORDS:** Spatiotemporal data; Gravity center; Standard Deviational Ellipse; Point of Interests; Visualization; Web Mapping

## ABSTRACT:

Analysing spatiotemporal distribution patterns and its dynamics of different industries can help us learn the macro-level developing trends of those industries, and in turn provides references for industrial spatial planning. However, the analysis process is challenging task which requires an easy-to-understand information presentation mechanism and a powerful computational technology to support the visual analytics of big data on the fly. Due to this reason, this research proposes a web-based framework to enable such a visual analytics requirement. The framework uses standard deviational ellipse (SDE) and shifting route of gravity centers to show the spatial distribution and yearly developing trends of different enterprise types according to their industry categories. The calculation of gravity centers and ellipses is paralleled using Apache Spark to accelerate the processing. In the experiments, we use the enterprise registration dataset in Mainland China from year 1960 to 2015 that contains fine-grain location information (i.e., coordinates of each individual enterprise) to demonstrate the feasibility of this framework. The experiment result shows that the developed visual analytics method is helpful to understand the multi-level patterns and developing trends of different industries in China. Moreover, the proposed framework can be used to analyse any nature and social spatiotemporal point process with large data volume, such as crime and disease.

## 1. INTRODUCTION

The spatial pattern of one industry may influence the development of enterprises of that industry and indicate the present focus of it (Higano, 2004), which is important for enterprises or government sectors to get an easy access to a clear demonstration of enterprises distribution. Meanwhile, to analyse and predict the trend of development of an industry, the research of its spatial pattern variation over time is also indispensable.

There already have been many researches conducted on the visualization and analysis of the spatiotemporal pattern of different datasets. A generalized space-time paths (GSTP) approach (Shaw et al., 2008) has been used to provide useful analysis environment for researchers who are interested in the individuals' immigration. With the fast development of Internet, new space-time visualization methods which concern relevant fields have been developed. Wallner et al. (2012) developed a system that uses clusters of nodes, pie-chart and coloured sections to assist the analysis of game players' behaviour. Ertl et al., (2012) uses seasonal trend decomposition based on locally-weighted regression (STL) to extract space-time information from social media. Although the spatiotemporal visualization has been used in many fields, the application in the field of space-time patterns of industries is still limited.

To offer as much information as possible and to keep the whole view clear at the same time, interactive visualization approaches and animation have been used to display maps selectively. Many novel approaches have been proposed, such as the shifting routes of gravity centers (Feng & Huang, 2006), spirals (Hewagamage et al., 1999), and a 3D timeline view called GeoTime (Kapler & Wright, 2005). Those approaches can help researchers effectively obtain information from huge quantities of data, but they are still relatively hard to understand for users without GIS knowledge. Meanwhile, many existing visualization systems are designed to present information in fixed spatial scale and attribute fields, users are not allowed to select spatial units, map levels and observation variable freely. Moreover, most of those systems are desktop-based which may increase user burden and lower the accessibility, as they have to be installed and configured appropriately on users' computers. In comparison, a web-based system may reduce the user burden in software installation and client-side computing resource occupation, as well as improving the accessibility to the public. To partially address above issues, we come forward with a web-based visualization framework using the shifting route of gravity centers and standard deviational ellipse (SDE) (Lefever, 1926). This framework offers information of different levels and industries, and users can switch between different levels and industries easily through interactive functions. The shifting routes of gravity centers allow users to learn about the variation of one industry over time, while SDE, which is widely used to measure the spatial properties of a particular point distribution, can show us the trend and the range of spatial coverage of an industry in one year. Those two techniques are especially useful when points are differentially weighted and of large amount.

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-129-2017 | © Authors 2017. CC BY 4.0 License.

Analysing and visualizing micro-level enterprise registration data with large spatial coverage can be challenging problems due to the huge data size. Mainstream scientific data processing languages or tools like R, Python and MATLAB show limitation and low time efficiency while dealing with big data. To solve this problem, distributed computing framework like Apache Hadoop (White, 2012) or Spark are developed to deal with such huge amount of data (Chen et al., 2014). Apache Spark (Zaharia et al., 2010), a newly developed high-performance clustering computing framework, has been proposed and widely-used in resent years. It outperforms Hadoop and other framework ten more times when dealing with big data due to the in-memory computing mechanism and also more flexible and portable since more programming languages are supported. Based on this reason, we chose Spark to achieve parallel data processing in our framework.

The research reported here aims at introducing the design of such a web-based visualization framework that uses the shifting routes of gravity centers and standard deviational ellipses (SDE) to present the spatiotemporal distributions and their changing patterns of different industries. The Apache Spark framework is used in this framework to improve the processing speed. The proposed framework can help end-users to obtain the spatiotemporal dynamics easily and in a user-friendly fashion.

## 2. METHODOLOGY

The proposed web-based framework is composed of three layers, i.e., data layer, service layer and web visualization layer. The logic structure of this framework is illustrated in Figure 1.

a) Data layer: Data layer is for data management. There are two databases in this layer, the registration information database and the result database. The former contains the detailed information of all enterprise records, the latter stores the computation results i.e., the gravity centers and SDEs of different spatial units and industries.

b) Service layer: RESTful web services are in charge of the communication between visualization layer and data layer. The services query and access data from the data layer, and sends computation results to the visualization layer. Spark is used to conduct computing if the computation results is not pre-calculated. Meanwhile, both the raw enterprise registration data and the computing result are cached in Spark using DataFrame, an embedded data model of Spark, to enable the parallel and in-memory computing with the help of Spark SQL and Spark streaming.

c) Web Visualization layer: This layer works as the client-side and graphical user interface (GUI) of the framework. It contains two basic views: map view and line chart view, which provide user with interaction and visualization functions to present the shifting routes of gravity centers and SDEs with figures and animations. To obtain required data for visualization, this layer sends HTTP GET requests to the service layer implicitly according to the user's requirement.

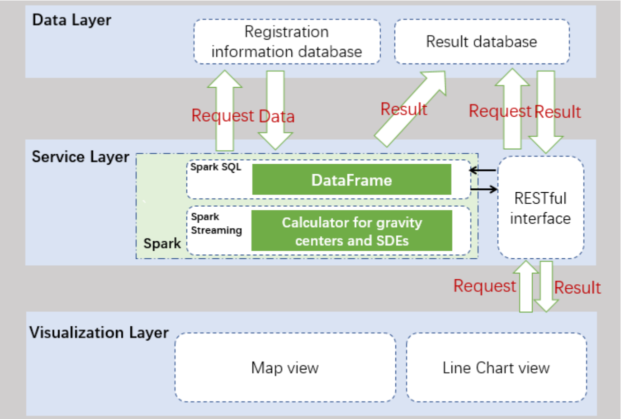


Figure 1. The structure of the proposed visualization framework

### 2.1 Data Layer

In this layer, we built up two databases, the registration information database and the result database. In the registration information database, each record records the registration information of an enterprise, and it contains many fields, including registered capital, registered year, industry code, address, coordinate and etc. The coordinate of each individual enterprise is obtained through public geocoding service (e.g., Baidu) by using its address information. According to the National Economic Industry Classification Standard (GB / T4754-1994) of China, all those industries are categorized into 16 main categories (China industry classification, 1994), as shown in Table 1.

| *Industry name* | *Industry code* |
|---|---|
| 1. agriculture, forestry, animal husbandry and fishery | AFAF |
| 2. extractive industry | EI |
| 3. manufacturing | M |
| 4. electricity, gas and water production and supply industry | EGWS |
| 5. engineering | E |
| 6. wholesale, retail trade and catering | WRC |
| 7. transportation, storage and communications industry | TSC |
| 8. finance and insurance | FI |
| 9. real estate | RE |
| 10. scientific research and comprehensive technical services | MI |
| 11. the geological prospecting industry, water conservancy | GP |
| 12. social services | SS |
| 13. education, culture and arts, film and television | ECAF |
| 14. health, sports and social welfare | HSSW |
| 15. state organs, political and social organizations | SPSO |
| 16. others | O |

Table 1. First-level industry categories (according to the National Economic Industry Classification Standard of China)

Since a SDE and a gravity center are related to each other in calculation, a computing result record contains the parameters of both SDE and gravity center in result database. When a query is triggered by specifying the industry code and spatial unit, the matched records of all the years will be retrieved. The fields and their meanings of the result table are showed in Table 2.

| Field name | Meaning |
|---|---|
| Industry | The industry that enterprise belonged to |
| Industry Code | The corresponding code of industry that enterprise belonged to |
| year | The year of information in this enterprise |
| Longitude | Longitude of the gravity center |
| Latitude | Latitude of the gravity center |
| spatialUnit | The spatial unit for the calculation, can be either the name of provinces or the name of nations. |
| theta | The intersection angle between the x-axis of the SDE and North |
| sigmaX | The standard deviations for the x-axis |
| sigmaY | The standard deviations for the y-axis |
| AreaID | ID representing the area this record belongs to |

Table 2. The table fields of the result table

The gravity centers and SDEs of industries within a fixed administrative division, i.e., a province or the whole nation can be pre-calculated and cached in the result database for further query. But if a user wants to upload new data to the framework or specify customized areas, calculation will be triggered in the Spark framework.

## 2.2 Service Layer

Due to the large amount of data, we implement the computation of SDEs and gravity centers in the service layer based on Spark framework using programming language Scala. The results for basic administrative division are stored in the data layer, and will be sent to the visualization layer through RESTful interface as JSON files. The computation methods for SDEs and shifting route of gravity centers are described in the section 2.2.1, while the Spark computing framework and the RESTful web services are introduced in section 2.2.2 and section 2.2.3 respectively.

**2.2.1 The computation methods:** We used the registration data from data layer to compute the shifting routes of gravity centers and SDEs. As enterprises of different business scales have different influence on the whole industry, we use weighted centers and weighted SDEs (Yuill, 1971) to show the pattern of different industries, and the registered capital is used to represent the weight of different enterprises. The computation of the gravity center of one region can be described as formula 1:

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i};\ \bar{Y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \qquad (1)$$

Where $(\bar{X}, \bar{Y})$ is the gravity center of the calculated spatial unit, and $(x_i, y_i)$ is the longitude and latitude of an enterprise within that spatial unit. The $w_1$, $w_2$, $w_3$… $w_n$ are registered capitals of those enterprises, which represent the scales of those companies.

To draw a SDE of the points located in a spatial unit, the gravity center $(\bar{X}, \bar{Y})$, the angle of rotation tanθ, the standard deviations for the x-axis and y-axis respectively, i.e., $\sigma_x$ and $\sigma_y$ are needed. The short half axe of a SDE shows the range of data and the long half axe shows the tendency of the data. When a SDE has been determined, its size can show the density of enterprises in that area, which is a necessary supplement to the gravity center.

The angle of rotation can be calculated as formula 2:

$$\tan\theta = \frac{\left(\sum_{i=1}^{n} w_i^2 \tilde{x}_i^2 - \sum_{i=1}^{n} w_i^2 \tilde{y}_i^2\right) + \sqrt{\left(\sum_{i=1}^{n} w_i^2 \tilde{x}_i^2 - \sum_{i=1}^{n} w_i^2 \tilde{y}_i^2\right)^2 + 4\left(\sum_{i=1}^{n} w_i^2 \tilde{x}_i \tilde{y}_i\right)^2}}{2\sum_{i=1}^{n} w_i^2 \tilde{x}_i \tilde{y}_i} \qquad (2)$$

where $\tilde{x}_i, \tilde{y}_i$ means the deviation of the x, y coordinates from the average center. Here we choose the gravity center as the average center.

The standard deviations of the x-axis and y-axis can be calculated as formula 3 and formula 4 respectively:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}\left(w_i \tilde{x}_i cos\theta - w_i \tilde{y}_i sin\theta\right)^2}{\sum_{i=1}^{n} w_i^2}} \qquad (3)$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}\left(w_i \tilde{y}_i cos\theta - w_i \tilde{x}_i sin\theta\right)^2}{\sum_{i=1}^{n} w_i^2}} \qquad (4)$$

**2.2.2 Spark frame:** In this web-based framework, the computation of SDEs and gravity centers is implemented in Spark. Two modules of Spark are mainly used, i.e., Spark SQL and Spark Streaming. The raw data and the final results are all stored using DataFrame, while some of the intermedia results are stored in RDDs. One DataFrame contains the data of one industry from one selected area of all years. This selected spatial unit can be either an administrative division or any customized area specified by end-user. For the former, the data of that administrative division can be selected directly by area field in the table, and for latter, latitude and longitude are used to filter records in the area.

Since we need to use the raw registration data from MySQL database, one module of Spark called Spark SQL (Armburst, et al, 2015) is used to query structured data from database into distributed memory using SQL statements directly. Spark SQL has two main advantages over previous systems: (1) it offers tighter connection between relational database and procedural processing, makes it easier for programming languages like Scala to process; (2) based on Spark framework, it costs far less time for every query. Spark SQL asks for the raw data of interested enterprises from the relational database, and the data acquired in this way is organized just like the way in the MySQL database, as it is stored in a DataFrame. DataFrame is a data model used in Spark, which is organized into named columns. This attribute makes it convenient for programmers to operate it with both MySQL statements and other popular programming languages (e.g., Scala, python). DataFrame can also easily be transformed into RDD, which allows us to implement complex computation in Spark Streaming module.

The Spark Streaming means that the input data stream is divided into several computation tasks, and they are put on the resource pool including massive computers, which can maximize the efficiency of data processing.

The computation of the parameters described in section 2.2.1 is implemented with Scala and the data flow is shown in Figure 2. The raw data will be firstly grouped by year, and then computed synchronously. As the value of gravity center is needed in the computation of SDE, the computing of SDE and gravity center is completed in the same process. The pseudocode of the computation of SDE and gravity center is illustrated in Table 3. The Scala-based code is distributed onto all the slaves in the Spark cluster and executed in parallel. The computation results of all the administrative divisions will be then write back into

MySQL database, and if the result is of a customized spatial area specified by users, it will be stored temporarily in one DataFrame and then sent to the visualization layer directly.

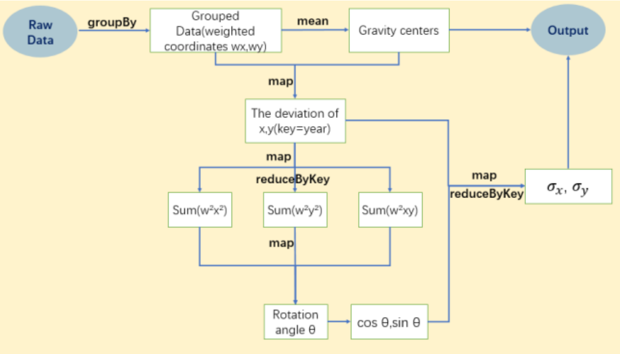


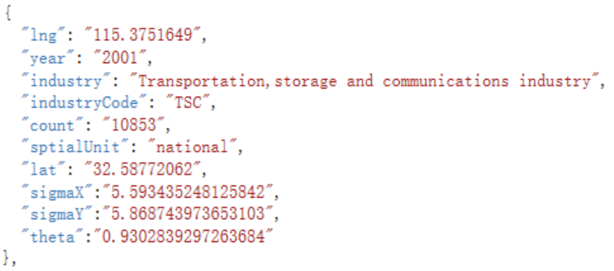
Figure 2. Data Flow of the computation of SDEs and gravity centers in Spark Streaming

| | | |
|---|---|---|
| 1 | conf 🡐 SparkConf() | //configuration of Spark |
| 2 | sc 🡐 SparkContext(conf) | |
| 3 | sqlContext🡐 SQLContext(sc) | // preparation for linking to relational database |
| 4 | need_infoDF🡐sqlContext.read.jdbc(url, TableName,properties) | //read data from relational database |
| 5 | gravityCenter🡐[need_infoDF.groupby(year).mean(latitude),need_infoDF.groupby(year).mean(longitude)] | //compute the gravity centers |
| 6 | infoWithCenter🡐need_infoDF.join(gravityCenter) | // using the year field to join two tables |
| 7 | intermediateResult🡐infoWithCenter.groupby(year). map{<br>x🡐ComputeDeviation(latitude)<br>y🡐ComputeDeviation (longtitude)<br>xSquare🡐x*x<br>ySquare🡐y*y<br>xy🡐x*y<br>} | //compute the deviation of the latitude and longitude of each enterprise from the gravity center in each year, and compute $x^2$, $y^2$, xy |
| 8 | SumOfIntermediateResult🡐[SumByYear(xy),<br>SumByYear(xSquare),<br>SumByYear(ySqaure)] | //sum up $x^2$, $y^2$, xy of all the enterprises in each year |
| 9 | tanTheta🡐SumOfIntermediateresult.groupby(year).map{} | //compute tan(θ) of data in each year |
| 10 | Theta🡐arctan(tanTheta) | |
| 11 | SigmaXandY🡐ComputeStandradDeviation(x, y,Theta) | //compute the standard deviation of x-axis and y-axis using x, y and theta |

Table 3. Pseudocode of the computation of SDE and gravity center

**2.2.3 RESTful interface:** RESTful web services are built to support the communication between client side and server side. It links the data layer to the visualization layer using HTTP GET operation. The raw data is selected by industry and spatial unit (can be either an administrative division or any specified area), and the result data is written into a JSON file to be used in the Visualization layer.

The result data is stored separately according to the administrative level of spatial unit, i.e., the national level or the provincial level. Figure 3 shows an example of the national level result, and the *spatialUnit* field identify the ID of the spatial unit in the database, i.e., *national*. For provincial level or customized spatial area, the value of this field is the name of one particular province and *customized* respectively.



```
{
  "lng": "115.3751649",
  "year": "2001",
  "industry": "Transportation, storage and communications industry",
  "industryCode": "TSC",
  "count": "10853",
  "sptialUnit": "national",
  "lat": "32.58772062",
  "sigmaX":"5.593435248125842",
  "sigmaY":"5.868743973653103",
  "theta":"0.9302839297263684"
},
```

Figure 3. An example of the calculation result of SDE and gravity center organized in JSON format

### 2.3 Visualization Layer

In order to show the spatial distribution variation of enterprises over time, web developing techniques are used. The visualization layer consists of two types of visual representation forms, map and line chart, as shown in Figure 4.

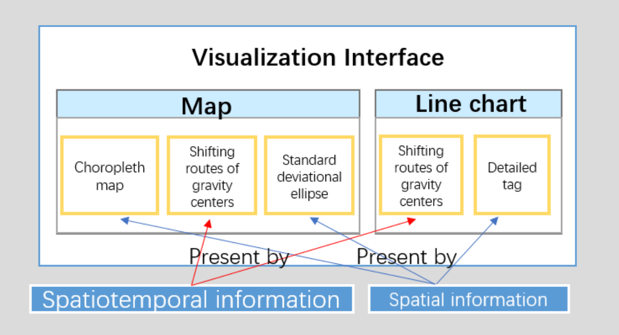

Visualization Interface
Map
Line chart
Choropleth map
Shifting routes of gravity centers
Standard deviational ellipse
Shifting routes of gravity centers
Detailed tag
Present by
Present by
Spatiotemporal information
Spatial information

Figure 4. The basic representation form of visualization layer

**2.3.1 Choropleth Map View:** The spatial distribution of gravity centers and SDEs are shown in this view for the selected spatial unit. The view has three main characteristics: (1) multi-scale choropleth maps are used to visualize variables; (2) interactive functions can avoid uninterested information from being shown and to highlight required information; (3) animation in this view helps to show and emphasize the temporal change of spatial features.

There are two map levels available in the choropleth maps, both the national map and provincial maps, based on the boundary of administrative divisions. The vector shapes representing SDEs and shifting routes of gravity centers are displayed upon the choropleth map. All the administrative divisions in the administrative divisions map are colored according to the overall number of enterprises of the chosen industry in that area to show development level of that industry.

Interactive functions are used to enable users to switch between different map levels, spatial unit, year and interested industry. User can click on the name tag of one industry to select industry. When clicking on a spatial unit on the choropleth map or a "*go back*" button, the observation spatial unit or map level will be changed. The SDE of one particular year appears when the marker representing a gravity center in a particular year is clicked.

The temporal pattern of one industry is illustrated with the help of animation. Markers representing gravity centers are linked by a polyline which indicates the shifting path, and those markers and line segments will appear year by year. Instead of using arrows to present the direction of the shifting direction of gravity center, we used animation to avoid the overlapping with centers and to attract users' attention.

**2.3.2 Line Chart view:** The line chart view is to offer detailed information of gravity centers. This view can be expanded to the full screen mode (Figure 5) which allows users to focus on it, and text tags attached to the marks of gravity centers can offer more detailed information.

Users can click on wherever on the chart to make this view zoom in or zoom out. This function allows users to focus on the shifting routes, and also helps them to have an overview of the map at the same time, as an eagle-eye map of the original choropleth map is put on the up-right corner of the screen too.

The line chart and the choropleth map is linked together, and the shifting routes of gravity centers appear synchronously on both of them. When users hover over a marker representing one year's gravity center, a tooltip offering information e.g., industry, longitude, latitude, and the overall number of enterprises of this industry that year, will appear on the page. As the gravity centers of some years didn't move much, the year labels of them may overlap with other figure elements or labels. To make the overall pattern of the shifting route line chart more clear, the framework offers functions to hide and show the year labels by clicking on the corresponding button (Figure 5).

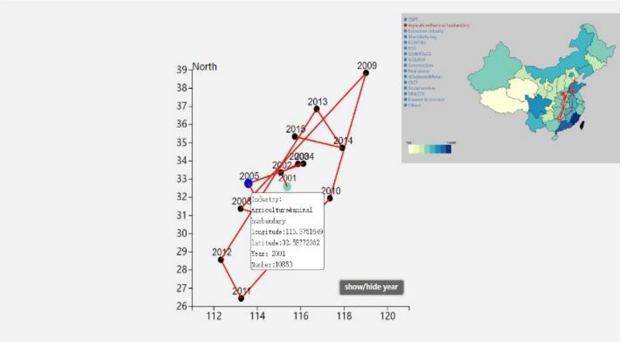

Industry:
Agriculture&animal husbandary
longitude:115.3751649
latitude:32.58772062
Year: 2001
Number:10853

Figure 5. An example of the line chart view that expended to full screen mode

## 3. EXPERIEMENT AND RESULT

### 3.1 Graphical user interface and functions of the framework

Using the framework proposed, we visualized the spatiotemporal distribution patterns of enterprises in China. We used the registration information of all the enterprises in Mainland China established between years 1950-2015 as the experiment data. The whole dataset has more than 16 million records accounted for 5 GB. The raw registration data is collected by local Administration in China and has been processed with machine learning methods and geocoding tools, to complete the industry category, administrative division and coordinates that each enterprise belongs to. (Li et al., 2017).

The main GUI of this framework is the map of China and a line chart. Users can choose one particular industry by clicking on its name (to save the space, some of them are abbreviations) to see its shifting route of gravity centers, which is shown in animation (Figure 6). Every red circle represents a gravity center of a year, and the red lines linking those circles show the shifting route. In this page, users can click on the province of interest to turn to the provincial page, or click on one circle to see the SDE of that year (Figure 7).

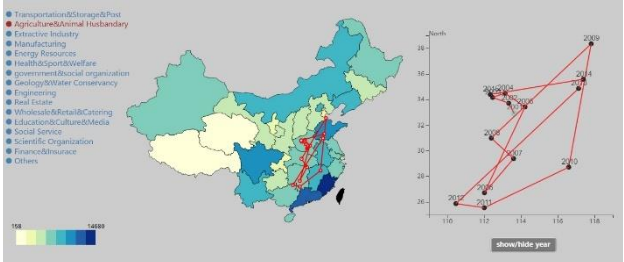


Figure 6. Graphical user interface of the nation level page

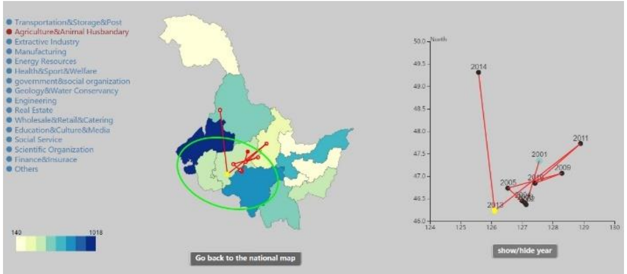


Figure 7. Graphical user interface of the province level page taking Heilongjiang province as an example

This framework provides users with necessary interactive functions, such as hovering over one area to get detailed information. Industry or gravity centers that have been chosen will be highlighted in different color to help identifying the choice of users. Main visualization functions of this framework are listed in Table 4.

| Function | Usage | View |
|---|---|---|
| Choose an industry | Click on the text representing industry or the circle before the text | Map |
| Go to the page of a province | Click on the area representing that province | Map |
| Go back to the national page | Click on the button | Map |
| Show the shifting routes of gravity centers | (Auto animation) After choosing an industry or change to anther page. | Map/Line Chart |
| Show the SDE of one year | Click on one circle representing that year | Map |
| Hide/Show the year text | Click on the button | Line Chart |
| Show detailed tooltip | Hover over one circle | Map/Line Chart |
| Change the color of selected industry or gravity center | (Auto) After choosing | Map/Line Chart |
| Color different areas according to the number of enterprises | (Auto) After choosing an industry or change to another page. | Map |

Table 4. Main functions of the visualization framework

https://doi.org/10.5194/isprs-archives-XLII-2-W7-129-2017 | © Authors 2017. CC BY 4.0 License.

### 3.2 Data Analysis

The visualization of shifting routes of gravity centers and SDEs can help users to explore the spatiotemporal distribution pattern of enterprises from different industries, or even to infer the trend of one industry.

From the shifting route of one industry, users can know which part of the study area has developed faster than others, and may contribute to the decision making of government sectors and enterprises. For example, from the shifting routes of some industries (e.g., wholesale, social services and scientific organizations) in Hubei, we can find that the gravity center kept shifting east progressively since 2006, and started to be distant from the geometric center of Hubei. The shifting route of wholesale industry in Hubei is shown in Figure 8. That phenomenon may be led by the fast development of Wuhan in these years, which locates in the eastern of Hubei. Due to this shifting trend, the government of Hubei may need to pay more attention to the development of the western part of this province to alleviate or erase the inequalities in regional developing.

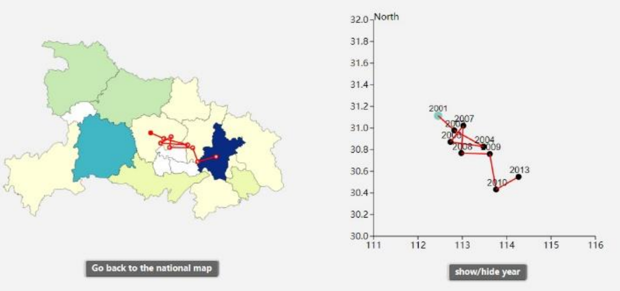
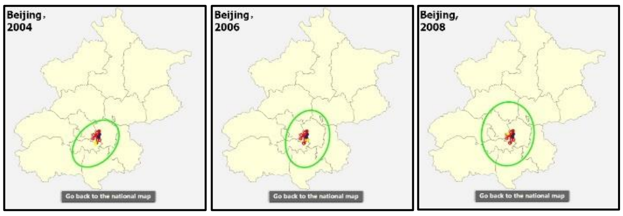


Figure 8. The shifting route of gravity centers of wholesale, retail trade and catering industry in Hubei, China

Standard deviational ellipses can show the density of enterprises and the direction of enterprises' distribution. For instance, by viewing the SDEs of the social service industry in Beijing from 2004 to 2008 (Figure 9), we know that in those years, the enterprises have distributed widely and isotropic as the size of this ellipse has grown larger, and the length difference between long axis and short axis has turned smaller.


Beijing，2004
Beijing，2006
Beijing,2008
Go back to the national map
Go back to the national map
Go back to the national map

（a） Year 2004 (b) Year 2006 (c) Year 2008

Figure 9. The standard deviational ellipses of social service industry in Beijing for year 2004, 2006 and 2008 respectively

### 3.3 Performance Comparison Experiments

To verify the performance and scalability of the Spark-based framework, we compare the computing time of the SDE and shifting route of gravity centers calculation algorithm written in Scala in terms of different number of CPU cores and different data sizes. The Spark cluster we used in experiments has 16 slave nodes. Each of the slave node has 8 virtual CPU cores of 2GHz, and 4 of them has been used in Spark computation. The memory of each node is 32GB. Figure 10 shows the computing time consumed by Spark to process different amounts of data, while Figure 11 illustrates the speed-up ratio of Spark with different numbers of cores compared with the time cost when using 4 CPU cores.

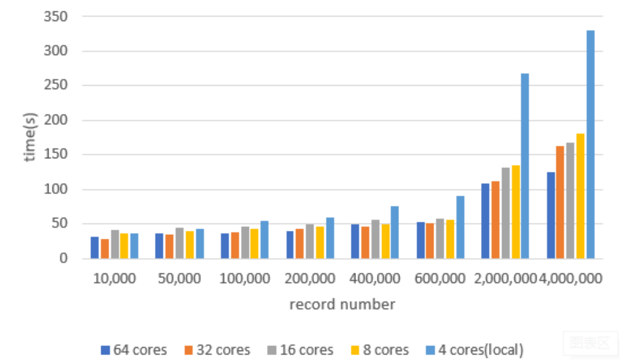


Figure 10. Computing time on different data size when using different number of CPU cores

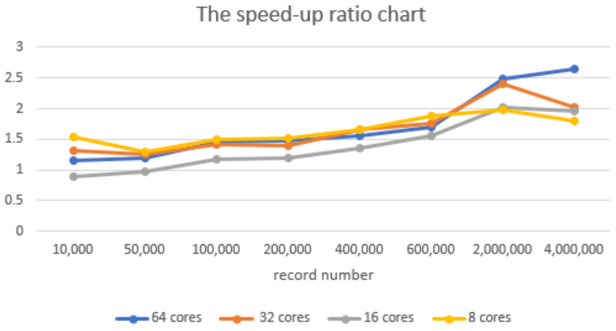


Figure 11. Speed-up ratios of computing time by using multi-cores compared with that using 4 CPU cores

Figure 10 shows that the proposed framework is scalable with the data size. The minimum time cost needed to process 600,000 records, is less than a minute. It means that it is possible for this framework to offer near real-time result. From Figure 11, we can find that the proposed framework is evidently more efficient when dealing with large amount of data by adopting Spark. The speed-up ratio varies in the data size and the CPU cores used. For smaller amount of data, 8 cores cluster has the best performance. While, when the data size getting larger and larger, more CPU cores show its advantage in computing. When data size is larger than 2 million points, the computing time of 64 cores begins to be less than that of 32 cores. The larger the data size is, the more computational acceleration. It shows that Spark is a powerful tool in big data processing and it is feasible to integrate Spark into our framework for the computation of SDEs and gravity centers.

## 4. CONCLUSION

In this paper, we designed and developed a web-based framework to visualize the spatiotemporal features of industry distribution using the shifting routes of the gravity centers and the SDEs. To demonstrate the feasibility and usage of this framework, we used the enterprise registration data of all industries in Mainland China registered between years 1950-2015 as a case study.

This framework can offer helpful information to the researchers and decision makers who are interested in the spatiotemporal distribution and its variation of industries with the help of the shifting route of gravity centers and SDE. Using both a multi-scale choropleth map and an animation-supported line chart, this framework presents the spatial distribution of those features and the temporal variation. Interactive functions allow the users to select information of interest easily.

Apache Spark is integrated in this framework to accelerate the computation, and experiments prove that the proposed framework and parallel algorithms are especially efficient when deal with large data size and can satisfy the computing and visualization need of real-time or near real-time. Apart from being applied to the analysis of industries, this framework can also help researchers to process different kind of datasets (such as population, number of crimes).

To offer more information about the development of different industries, additional functions can be added to this system in the future. For instance, comparison of SDEs of different years or comparison between different industries can be helpful. Also, it will be meaningful if users are able to input their own data through interaction in the visualization layer and generate shifting routes of gravity centers and SDEs of those data automatically.

## 5. ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundation of China (No. 41501434 and No. 41371372). We thank Fa Li and Yuan Wang from LIESMARS, Wuhan University for their considerable assistance with data processing and helpful suggestions.

## REFERENCES

Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., ... & Zaharia, M. 2015. Spark sql: Relational data processing in spark. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data.* ACM, pp. 1383-1394.

Chen, M., Mao, S., & Liu, Y. 2014. Big data: a survey. *Mobile Networks and Applications*, 19(2), pp. 171-209.

Ertl, T., Chae, J., Maciejewski, R., Bosch, H., Thom, D., & Jang, Y., et al. 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. *IEEE Conference on Visual Analytics Science and Technology*, 7, pp. 143-152.

Feng, Z. X., & Huang, J. S. 2006. Dynamic variation track and contrastive research of economic gravity centre and industrial gravity centre of china from 1978 to 2003. *Economic Geography*., 2, pp. 249-269.

Hewagamage, K. P., Hirakawa, M., & Ichikawa, T. 1999. Interactive Visualization of Spatiotemporal Patterns Using Spirals on a Geographical Map. *IEEE Symposium on Visual Languages*, pp. 296.

Higano, Y. 2004. The spatial economy: cities, regions, and international trade. *American Journal of Agricultural Economics*, 213(1), pp. 283-285.

Kapler, T., & Wright, W. 2004. GeoTime Information Visualization. *IEEE Symposium on Information Visualization*, 4, pp. 25-32.

Lefever, D. W. 1926. Measuring geographic concentration by means of the standard deviational ellipse. *American Journal of Sociology*, 32(1), pp. 88-94.

Li, F., Gui, Z., Wu, H., Gong, J., Wang, Y., Tian, S., Zhang, J., 2017. Big enterprise registration data imputation: Supporting spatiotemporal analysis of industries in China. *Computers, Environment and Urban Systems*. (Forthcoming)

Shaw, S. L., Yu, H., & Bombom, L. S. 2008. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS*, 12(4), pp. 425–441.

Wallner, nter, Kriglstein, & Simone. 2012. A spatiotemporal visualization approach for the analysis of gameplay data., *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, pp. 1115-1124.

White, T. 2012. Hadoop: The Definitive Guide. *O'Reilly Media, Inc.*

Yuill, R. S. 1971. The standard deviational ellipse; an updated tool for spatial description. *Geografiska Annaler. Series B, Human Geography*, 53(1), pp. 28.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. 2010. Spark: cluster computing with working sets. *Usenix Conference on Hot Topics in Cloud Computing*, 15, pp. 10-10.