The Quantitative Study of Geographical Elements
-- Establishment of Geographical Factors database

Dong Chun, Zhang Qingpu and Liu Jiping
Chinese Academy of Surveying , Mapping,
16 Beitaiping Rd., Haidian District, Beijing, China P.R
Email: dongchun@casm.ac.cn

Abstract In order to explore the correlation of geographical factors and economical factors by means of quantitative analyses, and provide the scientific basis for computer aided decision making of national economy, based on the theories of Data Mining, the concept of Geographical Factors Database is put forward, and the principles and methods for establishment of geographical factors database are studied. In this paper, the geographical factors are obtained from topographic database and DEM database at scale of 1:250 000 within the territory of Fujian Province. The fundamental geographical factors databases covering county administrative boundaries and 1km*1km grid are set up; the economical development factors database based at county level has been set up as well. Both geographical factors database and economical development factors database are able to offer sufficient conditions for quantitative analysis. Experiment results prove that Geographical Factors Databases have the important and basic roles in fields of data mining, DMKD, and regional sustainable development.

Keywords geographical element, geographical factors database, economical development factors database, digital mining, quantitative analysis

1. Introduction

Now, people have enough ability in producing and collecting data, so people can apply thousands databases easily in commerce management, government work, science research, and engineering exploitation etc, and more and more databases will apply in more fields. So, a new challenge how to find the knowledge in the boundless information ocean is put forward.

Spatial Decision Support System(SDSS) are becoming important tools for planning and decision making for environmental management(K.Taylor, G.Walker and D.Abel, 1999), in domains like land use planning, biodiversity preservation, catchment management, urban development, and the forestry and mining industries, decisions are made about issues with complex physical and social implications within complex organizational frameworks. In this setting, an SDSS has to combine spatially explicit observational data and simulation non-specialist decision makers and other stakeholders. Geographical Information System (GIS) enable a familiar interface paradigm for specification of decision scenarios and presentation of predicted outcomes; model-based simulation systems provide the means for scientific analysis of decision scenarios.

In the early stages of GIS development, only five functions such as entry / updating, data conversion, storage / organization, manipulation, and presentation / display were envisioned by GIS developers. A sixth required function became apparent as soon as the technology advanced to a degree where GIS became useful for empirical applications. According to ESRI's definition, in addition to others mentioned above, a complete GIS must provide spatial analysis function (ESRI, 1994). Indeed, it is the spatial analysis capability that differentiates GIS from desktop mapping software (Fotheringham, S and P.Rogerson, 1994, Yue-Hong Chou, 1997). Hypothetial distributions can be generated with specified factors, while descriptive, explanatory, and predictive models can be derived.

Analysis of area features is most common in quantitatively oriented spatial analyses. Area units of spatial analysis can be defined in different ways. The grid and polygon are common.

It's well-known that the developing of crop especially paddy is enable without water, so the distributing of water resource such as river, lake and so on is the sticking point if planting paddy enable in a special region. Bigger and more hospitals, schools locate in big cities than in countries in China. More than 85 percents of the information on government affair is related to the information on spatial position (Zhang Qingpu, 1999). So we conclude that much information on national economy and social development has certain correlation with different geographical factors in a rejoin. In order to explore the determinate correlation of geographical factors and economical factors by means of quantitative analyses, Zhang Jiaqin, Dong chun etc. put forward the concept of Geographical Factors databases and established the fundamental geographical factors databases covering county administrative boundaries and 1km * 1km grid within the territory of Fujian Province.

2. Definition and content of Geographical Factors database

Geographical factors are the essential components in substance and energy of geographical environment. They are self-existent, having different properties, but obeying the holistic evolving rules (Niu Wenyuan, 1991). Despite the Geographical factors are reciprocation, interplay and coadjustment, no factor can be replaced. In this paper, geographical factors databases are defined as databases that stored the quantitative values of multifarious geographical factors in each geographical cell.

Every geographical cell represents the lowest area of the geographical space. Here geographical cell shapes administrative boundary or grid cell. We can establish district –based geographical factors database covering county administrative boundaries, province administrative boundaries or village. and grid-based geographical factors database covering arbitrary grids as well. Considering the information we had obtained, we established the fundamental geographical factors databases covering county administrative boundaries and 1km * 1km grid cells within the territory of Fujian Province.

Figure 1 shows the Possible Categories of geographical factors databases.

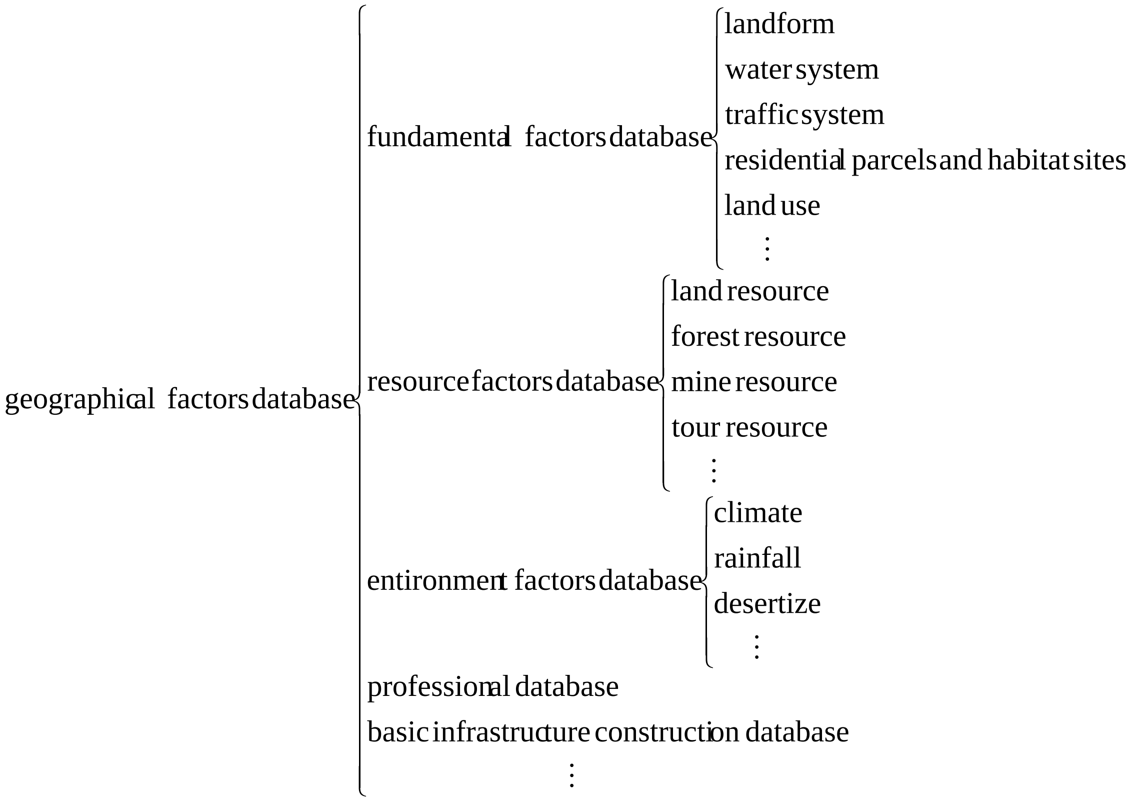

geographical factors database
fundamental factors database
landform
water system
traffic system
residential parcels and habitat sites
land use
⋮
resource factors database
land resource
forest resource
mine resource
tour resource
⋮
entironment factors database
climate
rainfall
desertize
⋮
professional database
basic infrastructure construction database
⋮

Figure 1. The Possible Categories of Geographical Factors Database

3. Establishment of Geographical Factors database

3.1 *The Data Structure of Spatial Data*

We know that spatial data can be organized in either a raster or vector data structure. In the raster structure, spatial features are organized in a regularly spaced coordinate system. The DEM database at scale of 1:250 000 we applied was organized in a raster structure. Every points represent the altitude of hypsography or the depth of seabed on the site. The size of cells is 100m * 100m. Alternatively, in the vector structure, spatial features are defined and organized by combinations of vectors. ESRI’s ARC/INFO and PC ARC/INFO is example of vector-based GISs. The data structure adopted in a GIS determines how different pieces of information are organized and processed, and thus also determines the properties of the GIS for performing spatial analyses. (Burrough, 1986; Clarke, 1995).

The topographic database at scale of 1:250 000 we applied used the structure of LIBRARY module, ARC/INFO. And the unit of Transverse Axis is TILE and Lengthways Axis is LAYER. There are 816 tiles covering the territory of China, and 18 covering the territory of Fujian Province. And we can obtain geographical factors from following layers:

HYDNT（water system represented as points and polygons）、HYDLK（water system of points and lines）、RESPT（residential area represented as points）、RESPY（residential area represented as polygons）、RAILK（railways represented as points and lines）、ROALK（roads represented as points and lines）、BOUNT（boundaries represented as lines）etc.

In the vector data structure, every spatial feature is represented by a set of vectors. In mathematical terms, a vector is specified by a starting point (with given x and y coordinates), a direction (i.e., an angle toward east, west, north, south or some indeterminate direction), and length.

Points are the simplest form of spatial features. A point feature is represented by a "degenerate" vector with both direction and length equal to zero. In this case, the point feature is not associated with a valid measure of area.

Line features can be treated as combinations of ordered point features. Every line feature is one-dimensional and represent both position and direction. Length is a significant measurement of line features. Although linear features actually occupy two-dimensional space on maps, their width is not considered in cartographic renderings. For instance, roads and rivers are commonly represented as line features. This representation is meaningful only in terms of length, although their actual measurements are both length and width. Therefor, the width of a vector is not a valid measurement, and only the length of connected vectors is meaningful.

Polygon features are the most complicated among the three types of spatial features. Polygons are defined by a series of line features delineating boundaries. A polygon feature is represented by a series of vectors that form an enclosed area. In other words, polygon features are two-dimensional and represent both position and area. And the area of a polygon is a valid measurement.

3.2 *Obtaining Geographical Factors*

ESRI's ARC/INFO is based on the arc-node data model and organized spatial feature data in separate, relation data files[ ]. The following illustration (Figure 2) shows an example of a map comprised of four regions represented by polygons (101,102,103, and 104). The polygons are delineated by seven arcs (33 to 39). Note that every arc has two nodes (start and end). In the arc-node data model, arcs form the most basic units on a map, every arc consists of two nodes, a start node and an end node. Between the nodes an arc could have zero or any number of vertices. And a point feature can be treated as a degenerate line feature with the start node overlapping the end node and no vertices in between. Each point feature can be represented by one node because start and end nodes are identical. Each polygon feature in the previous illustration, however, is treated as a series of connecting arcs that delineate the boundaries of the polygon.

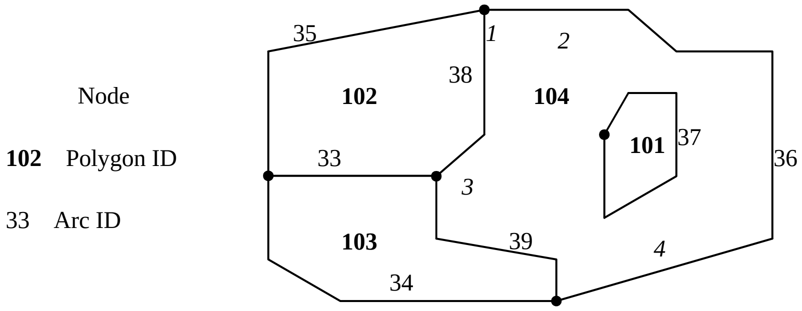


Figure2 A simple polygon coverage of the arc-node data model

Appearing below are the three most important attribute tables for obtaining geographical table (PAT) (table1) and the arc attribute table (AAT) (table2). Considering the particularity of point attribute, a new record should be add to the PAT (point attribute table) , see table 3.

Table1 **PAT**(polygon attribute table)

#-ID	*Poly-ID*	*Perimeter*	*Area*
1	0	8.418	-4.056
2	104	8.596	2.078
3	102	4.296	1.144
4	101	2.233	0.301
5	103	4.325	0.983

Table2 **AAT**(arc attribute table)

#-ID	*Arc-ID*	*F-node*	*T-node*	*l-poly*	*R-poly*	*Length*
1	38	3	1	3	2	1.151
2	33	4	3	3	5	1.040
3	35	4	1	1	3	2.105
4	37	2	2	2	4	2.233
5	36	1	5	1	2	4.120
6	39	5	3	5	2	1.093
7	34	4	5	5	1	0.983

Table3 **PAT**(point attribute table)

#-ID	*Poly-ID*	*Perimeter*	*Area*	*Number of points*
1	1	0.000	0.000	1
2	2	0.000	0.000	1
3	3	0.000	0.000	1
4	4	0.000	0.000	1

Apparently, the quantitative value in any geographical cell is the accumulative total of length(arc), number of points(point), or area(polygon). In this way, we got more than one hundred factors from topographic database and DEM database at scale of 1:250 000 within the territory of Fujian Province.(Figure 3).

On the other hand, three factors was obtained from the DEM database at scale of 1:250 000, they are gradient, direction of slope and uneven degree of the earth’s surface. The calculation formula we applied can be fined in reference .

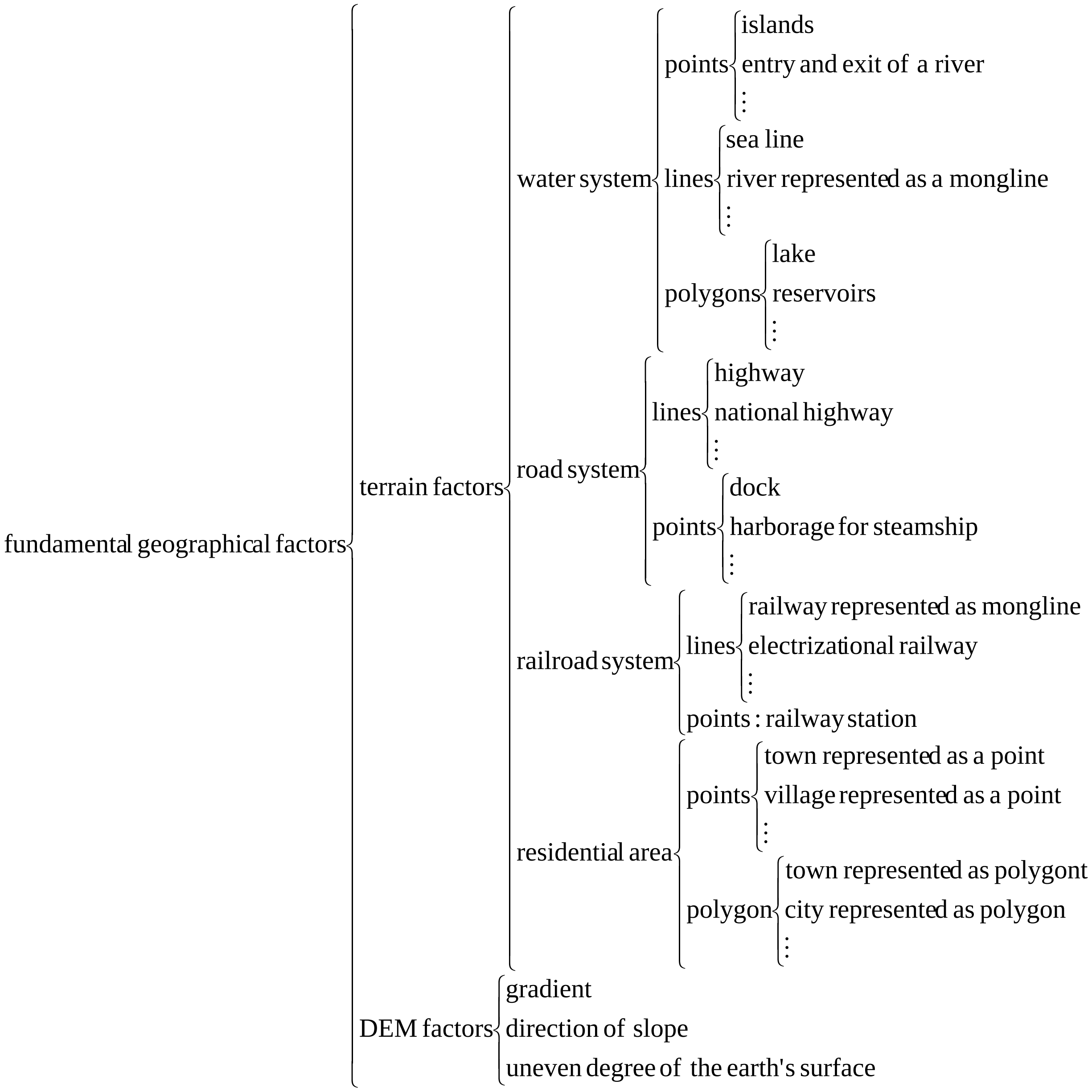


Figure 3. The Contents of a Geographical Factors Database

4. Results

4.1 *Geographical Factors Databases and Economical Development Factors Database*

Table 4 and table 5 shows the structure of the district–based(County Administrative Boundaries) geographical Factors and grid–based(km*1km) geographical Factors in Access databases.

Table 4 lists the geographical Factors databases based on County Administrative

-

Boundaries. The first column lists Political Area Code (PAC), and the second column lists Political Area Name, geographical factors are listed from the third column to the end.

Table 4. The Structure of County Administrative Boundaries Based Geographical Factors Database

Political Area Code	Political Area Name	perennial rivers(mongline, meter)	Perennial lake(square meter)	Railway station(Entries)	…
350100	Fuzhou City	686242.49	837952	3	…
350121	Minhou County	1465993.249	736115	3	…
350122	Lianjiang County	868302.083	261215.25	2	…
350123	Luoyuan County	728541.26	41309	0	…
350124	Minqing County	885641.753	46718	0	…
⋮	⋮	⋮	⋮	⋮	

Table 5 lists the geographical factors databases based on grid. The first column lists serial numbers of geographical cells, there are two parts consists of ID, the first five digit and the final four digit, the former represents the horizontal serial numbers of cell, and the latter represents the vertical serial numbers of cell with origin as (-2700 km，1875 km) in Albers projection. Geographical factors are listed from the second column to the end.

Table 5. The Structure of 1km*1km grid Based Geographical Factors Database

ID	perennial rivers (mongline, meter)	perennial rivers (crewel, meter)	Season rivers (mongline, meter)	Season rivers(crewel, meter)	…
303404270	0	0	0	0	…
303404271	0	0	0	0	…
303404272	450.253	0	0	0	…
303404273	0	0	0	0	…
303404274	420.682	0	0	0	…
303404275	0	75.5	0	0	…
303404276	394.665	790.581	161.379	0	…
⋮	⋮	⋮	⋮	⋮	…

In order to analysis the correlation between geographical factors and economical development factors, Economical development factors database based on county administrative boundaries was built as well, table 6 is the structure.

Table 6. The Structure of County Administrative Boundaries Based Economical Development Factors Database

Political Area Code	Political Area Name	Amount of Regional Finance(ten thousands yuan)	Highroad length(meters)	Medicinal professional(points)	…
350100	Fuzhou City	240490	429.4	16683	…
350121	Minhou County	13744	744.7	1082	…
350122	Lianjiang County	13189	500.32	1072	…
350123	Luoyuan County	5891	440.48	610	…
350124	Minqing County	8871	659.64	826	…
350125	Yongtai County	4882	521.62	785	…
350126	Changle City	18457	358.53	1302	…
⋮	⋮	⋮	⋮	⋮	

4.2 *The relativity between Geographical Factors and Economical Development Factors*

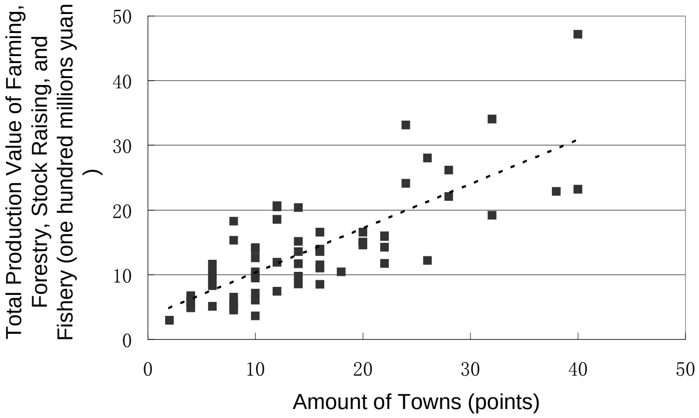
When the study variables are measured quantificationally, either the spatial correlation between different spatial feature types, or the correlation between geographical factors and economical development factors is showed in more appropriately tested by Pearson’s correlation coefficient or regression analysis.

Table 7 Several Geographical Factors and Economical Development Factors Applied in Experiment

Amount of Towns (points)	Total Production Value of Farming, Forestry, Stock Raising, and Fishery (One Hundred Millions)	Total Population (Ten Thousands)	Amount of Towns (points)	Total Production Value of Farming, Forestry, Stock Raising, and Fishery (One Hundred Millions)	Total Population (Ten Thousands)
2	2.95	9.9	14	13.56	34.79
4	5.89	15.64	14	9.76	39.27
4	5.75	15.02	14	20.35	55.12
4	6.73	12.54	14	11.69	44.91
4	6.48	21.16	14	15.17	33.74
4	4.95	15.1	16	13.94	47.43
4	4.87	26.43	16	13.55	47.28
6	9.01	48.35	16	8.5	27.19
6	9.54	36.06	16	16.56	53.04
6	11.63	24.47	16	11	42.7
6	5.09	18.05	16	11.52	29.93
6	10.69	34.73	16	13.64	33.34
6	8.31	18.56	18	10.45	29.95
8	15.35	62.71	20	14.95	34.45
8	6.54	16.78	20	15.06	54.24
8	4.56	11.6	20	14.58	40.24

8	5.06	14.44	20	16.55	50.27
8	5.63	21.03	20	14.97	141.68
8	18.3	37.56	22	14.25	95.96
10	13.9	19.89	22	11.73	48.28
10	10.44	35.83	22	15.98	58.29
10	7.15	15.86	22	15.9	103.11
10	12.58	39.69	24	24.1	125.73
10	9.94	23.45	24	33.15	61.1
10	9.45	32.6	26	28.01	76.56
10	6.1	29.97	26	12.2	52.93
10	3.66	18.6	28	26.14	77.95
10	14.18	25.19	28	22.07	67.17
12	20.46	60.52	32	19.16	99.46
12	11.91	45.92	32	34.04	124.67
12	18.55	41.57	38	22.86	147.15
12	20.64	50.01	40	23.19	158.69
12	7.45	24.09	40	47.17	116.89

The relationship between geographical factors and economical development factors can be examined from the next two illustrations (Figure 4), which is derived from the preceding table 7. In the first diagram, the vertical axis represents total production value of farming, forestry, stock raising, and fishery in each county, the horizontal axis represents amount of towns in each county (1996). And in the second diagram, the vertical axis represents crop yield, the horizontal axis represents the length of perennial river(mongline) in each county (1996). Every dot in the diagram represents a county plotted according to its corresponding values. The diagram shows a corresponding relationship between these two variables—greater amount of towns implies greater total population and greater total production value of farming, forestry, stock raising, and fishery.



-

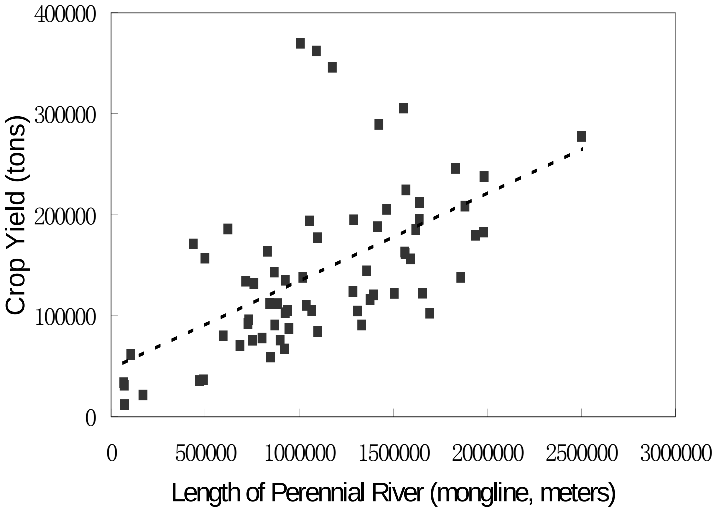


Figure 4 Two Scatters Chart for a Geographical Factor and a Economical Development Factor

To confirm this relationship and make it precise, Pearson’s correlation coefficient can be used. The correlation coefficient (*r*) between two variables, *x* and *y*, is defined as follows:

$$r = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n s_x s_y}$$

Where n is the number of counties in Fujian Province and n = 67 in previous example), and s_x and s_y denote the standard deviation of variables *x* and *y*, respectively.

The correlation coefficient (*r*) indicates the degree and direction of the relationship between two variables. The values of range between –1 to 1. A positive *r* value indicates that there is a positive (direct) correlation between the two variables, that is, a large value of one implies a large value of the other. At the extreme, *r* = 1, the two variables are perfectly correlated and their distribution patterns must be identical. A negative *r* value implies a negative (indirect) correlation; the variables are inversely related to each other. A higher value of one variables implies a lower value of the other. If the *r* value is not significantly different from 0, there is no correlation between the variables are considered independent of each other.

In the example, the positive correlation between total production value of farming, forestry, stock raising, and fishery and amount of towns as well as total population and amount of towns in each county is obvious. The computed *r* equals 0.786 and 0.573, indicated a significant positive correlation between the two groups variables.

5. discussion and conclusion

-

In this paper, a method testing the relationship between geographical factors and economical development factors is put forward. We know agriculture is very important in Fujian province, for which takes the very advantage of warm climate, soil adaptability, and important abundant water resource etc. In other words, all of these conditions are favorable to the production of crop especially paddy. Furthermore, rivers in Fujian province which acting leading actor to irrigated farming.

Upon that we can draw a conclusion that it’s doable and useful for us to mine information in geographical factors databases and economical development factors databases (Data Mining). The establishing of geographical factors databases will apply a feasible, scientific and convincing support technically in the way of discovering knowledge from database, developing economy sustainedly in a region.

At the same time, it’s a neglect in this paper that we ignoring the effect among different geographical factors, and the effect could bring into fatal effect to the result we gained. Therefore some complicated mathematical models, expert system, and some knowledge and method of other subjects are integrant parts for data mining(DM) and knowledge discovering from databases(KDD).

acknowledgements

The work done by two anonymous referees is greatest acknowledged; their comments and suggestions significantly improved this manuscript. Prof. Zhang Jiaqing, vice Prof. Wang Liang provided very helpful comments on an earlier draft of the paper. The authors are grateful for the many useful suggestions made by the anonymous reviewers.

References

Dong chun. 2000. *A Research for the Establishment and Application of Geographical Factors Databases.* Thesis for Master's Degree. Beijing: Chinese Academe of Surveying and Mapping.

Dong chun, Zhang qingpu, and Zhang jiaqin etc. 2000. *A Discuss on the Establishment and Application of Geographical Factors Databases.* Remote Sensing Information.

Dong chun, Wu xizhi, and Chen bo. 2000. Study on Application of PLS in Correlation Analysis Between Geographical Parameters and Economical Development Parameters. Science Surveying and Mapping.

K.Taylor, G.Walker and D.Abel,1999, *A framework for model integration spatial decision support system*. International Journal of Geographical Information System,13, 533-555

Burrough, P.A. 1986. *Principles of Geographical Information Systems for Resources*

Assessment. Oxford: Clarendon Press.

Clarke, K.C. 1995. *Analytical and Computer Cartography*. Englewood Cliffs, New Jersey: Prentice Hall.

Environmental System Research Institue. 1994. *Understanding GIS: The ARC/INFO Method, Version 7 for UNIXand OpenVMS*. Redlands, California: ESRI.

Fotheringham, S. And P. Rogerson (eds.). 1994. *Spatial Analysis and GIS*. London: Taylor &Francis.

Chou, Y. H. 1997. *Exploring Spatial Analysis in Geographic Information Systems*. Santa Fe: Onword Press.

Zhang Qingpu. 1999. *The Construct Mode and Run Mechanism of Government GIS*. Beijing: The Corpus for Celebrating the Founding of CASM Forty Years Ago.

A dictionary of modern geography. 1990. The Commercial Press . p29

Batty M，Xie Y. 1994,8(3). Modelling Inside GIS：Part 1. Model Structure，*Exploratory Spatial Data Analysis and Aggregation*，Int.J.GIS.

Huang Yingyuan, Tang qin etc. 1991. *The Conspectus of Geographical Informaion System*. Beijing: Higher Education Publishing House. China Higher Education Press.