CHINESE GEOGRAPHICAL SCIENCE
Volume 14, Number 4, pp. 337−342, 2004
Science Press, Beijing, China

A SPATIAL CLUSTER METHOD SUPPORTED BY GIS FOR URBAN-SUBURBAN-RURAL CLASSIFICATION

ZHOU De-min[1], XU Jian-chun [2], John RADKE[2], MU Lan[2]
(*1. Northeast Institute of Geography and Agricultural Ecology, Chinese Academy of Sciences, Changchun 130012, P. R. China; 2. Geographical Information Science Center, University of California, Berkeley, CA 94720, U.S.A.*)

ABSTRACT: This study was undertaken to construct a preliminary spatial analysis method for building an urban-suburban-rural category in the specific sample area of central California and providing distribution characteristics in each category, based on which, some further studies such as regional manners of residential wood burning emission (PM2.5, the term used for a mixture of solid particles and liquid droplets found in the air, refers to particulate matter that is 2.5μm or smaller in size) could be carried out for the project of residential wood combustion. Demographic and infrastructure data with spatial characteristics were processed by integrating both Geographic Information System (GIS) and statistics method (Cluster Analysis), and then output to a category map as the result. It approached the quantitative and multi-variables description on the major characteristics variations among the urban, suburban and rural; and perfected the TIGER's urban-rural classification scheme by adding suburban category. Based on the free public GIS data, the spatial analysis method provides an easy and ideal tool for geographic researchers, environmental planners, urban/regional planners and administrators to delineate different categories of regional function on the specific locations and dig out spatial distribution information they wanted. Furthermore, it allows for future adjustment on some parameters as the spatial analysis method is implemented in the different regions or various eco-social models.
KEY WORDS: GIS; cluster analysis; PM2.5; census tract; urban-suburban-rural classification

CLC number: P208 Document code: A Article ID: 1002-0063(2004)04-0337-06

1 INTRODUCTION

The project of residential wood combustion aimed to develop a statewide base model for classifying the location of residential households and estimating the potential residential wood burning sites in temporal and spatial manner. The study was to assist air quality management to better understand how residential wood burning affected air quality and how its impacts on air quality could be minimized. Before carrying out the further study on specific regional characteristics of residential wood burning, researchers had to figure out the spatial boundaries of the urban, suburban and rural, because behaviors of residential wood burning in various regional categories were significantly different.

The existing methods for urban-suburban-rural classification are based on the statistical data of population density, regularly, at the different levels of administrative divisions. Among them, the TIGER's (Topologically Integrated Geographic Encoding and Referencing) scheme of urban-rural classification is typical. The area of population density above 1000 per square mile (386.27/km^2) will be defined as the urban area; the other will be the rural area (U.S.A. Census, 2000). However, this scheme will be challenged by fellow two problems.

First, TIGER′s scheme is an urban-rural two level classification without suburban category. It is obviously unreasonable if the suburb will be excluded while the suburbanization process (sprawl, the regional development process of bringing the city into the rural fringe (ANDRE, 2000; MARET and DAKAN, 2003), or exurbality, the development of very low density land found around small towns and major metropolitan areas alike (NELSON, 1992)) has become one of the most noticeable aspects in the regional development process and its impact can be perceived everywhere. Second, as the juncture between the urban and rural, both configuration and distribution of suburban area are very complex. Ac-

Received date: 2004-04-21
Foundation item: Under the auspices of the research contract from California Air Resources Board (ARB) and the Talented Foundation of Northeast Institute of Geography and Agricultural Ecology, Chinese Academy of Sciences (No. c08y17)
Biography: ZHOU De-min (1967−), male, a native of Dexing of Jiangxi Province, associate professor, specialized in Geographic Information Science and Environmental Sciences. E-mail: zhoudemin@neigae.ac.cn

tually, it is very difficult to implement the classification between the urban and suburban, or suburban and rural if we use the only criterion of population density (HATHOUT, 2002).

From a wide scope of regional occupational, socio-cultural and life-style perspective, and residential economic performance, this preliminary methodology was explored for undertaking multivariate analysis instead of the single scope of population density, so that the suburban as a specific category could be approached in our study.

2 STUDY AREA AND DATA SOURCES

As a large state in the United States, the California State itself contains rich diversity of both regional distribution and eco-social entities. The researchers observed the rich regional characteristics, and tried to represent the various characteristics in the study area.

The study area included 28 counties in the central California State (Fig. 1). This region, with more than 9×10^6 population in a large area of 76 238km^2, represented a good cross-section of California as it was characterized by multiple patterns of regional development, including the metropolitan area such as bay area and the capital of the state, Sacramento; it contained a suitable proportion of coasts and inlands, the metropolitan area in the western coastal area and small towns and countryside in the eastern inland; and it represented a wide distribution of the state's demographics.

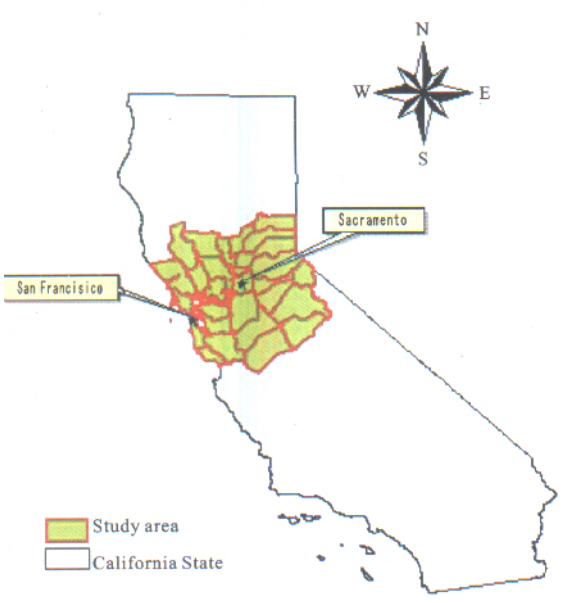

Fig. 1 Location sketch of study area in California

The census data is a digital database system of geographic features, such as roads, railroads, rivers, lakes, census statistical boundaries, etc. covering the entire United States. The database contains rich information about features such as location in latitude and longitude, name, type of various geographic features, address ranges for most streets, the geographic relationship to other features, and the other related (U.S.A. Census, 2000). They are the public product from the Census Bureau′s TIGER .

Census tracts are small, relatively permanent statistical subdivisions of a county, defined by local participants as part of the U.S.A. Census Bureau′s Participant Statistical Areas Program. Other data sources include 2000 TIGER Street as a key aspect of the infrastructure; also, coastline and water body data (U.S.A. Census, 1995) will be presented for area calculation.

3 METHOD

3.1 Data Preprocessing

As the first step for modeling, initial census tract data with Geographic Projection were transformed into Alberts Projection, but its datum was exchanged from nad27 to nad83, so that the projection bias could be prevented for area calculation. The coastline and water body were clipped out from the census tracts as the next step in case of the bias from calculation of the density data based on lands, which consists of water body occupation or coast occupation.

Data preprocessing produced some very small polygons, these small polygons usually had very high-density or ratio values, and most of them were located along with the boundary of tracts. We need to rebuild our polygon by using eliminate function in ArcGIS 8.3, it would be logically dissolved 199 polygons with area smaller than 0.1km^2 into mostly nearby polygons. Now, we get a 28 counties polygon, with 1926 tracts as our study area.

3.2 Selection of Spatial Variables

Common variables referred for regional planning usually include: population, housing, transportation, economy, utilities, facilities and services, health and safety, land, urban development, plan and policies. Each variable includes some sub-variables, for example, the housing variable usually includes number of dwelling units, dwelling type, quality, homeless population, housing vacancies, and supply and demand (ZORICA *et al*., 2004).

From 198 spatial variables (attributes ranging from geographic identifiers to population, housing, race/ethnicity, family, income, education, commuting, etc.) available

in the demographics of census tract, 16 variables (Table 1) were finally chosen as potential variables for undertaking the enlarged variation of urban-suburban-rural classification. From occupation density, infrastructure level, educational level, housing structure, residential income structure, life style to relative social and cultural contradictions, these variables represented the various aspects of regional characteristics. Among all these 16 variables, some were eliminated for data definition biasing from our conception, and some high correlated variables were eliminated in case of some variables over fit in the cluster process. The final 7 variables (Table 2) for cluster analysis would critically classify the urban, suburban and rural, and we tested correlation of these 7 variables, which were all not in high correlation (Table 3).

Table 1 Sixteen potential variables for cluster analysis

Variable	Variable
1. Road density	2. Housing density
3. Rate of home sale	4. Public water/private water
5. Gas consuming	6. Public sewer/private sewer
7. Farm income	8. Size zip code (mail)
9.Wealthy in rural area	10. Agriculture sales
11. School density	12. Public transportation
13. Work in-city or ex-city	14. Apartment duplex
15. Rent occupied or owned	16. Population density

3.3 Cluster Analysis

By analyzing the spatial characteristics (demographics) of each small spatial unit (census tract), we could divide a region into various categories, which are clustered by some adjacently small units with most similar regional characteristics (ALAN and VLADIMIR, 1998). The classification map of the urban, suburban and rural were finally embedded within the GIS as the result of quantitative multivariate analysis, which would be used to analyse and predict residential wood burning behaviors of specific social groups that geographically clustered. The methodology was supported by the theory that 'Neighborhoods with similar demographic characteristics have similar tastes, lifestyles, and consumer behavior. These behaviors are measurable and predictable and, therefore, can be targeted" (LAWRENCE, 2003)

The goal to classify regional categories could be approached by clustering all the small units, which composed the whole region, into the different categories, since every unit had both a definite boundary and specific spatial distribution. Based on the hierarchical spatial distribution of each unit, we could build up the categories by clustering these small units (WANG *et al*., 1998).

Table 2 Final 7 variables for cluster analysis

Variable	Abbreviation	Definition	Data source (Demographics)
Population density	POP_DEN	Total population/area	Persons, area_km^2
Ratio of public water consuming	R_PUBWA	Public water consuming/household	Pubwa, household
Road density	ROAD_DEN	Aggregate road length/area	Length (TIGER 2000 STREET), area_km^2
Ratio of bottle gas consuming	R_BOTGA	Bottle gas consuming/household	Botl_gas, houshold
Ratio of renters	R_RNT	Renter occupied units/household	Rent_occ, houshold
Ratio of farm income	R_FARMI	Farm income in household	Farm_inc
Rent price	RNT_MEDI	Median rent	Rnt_medi

Table 3 Correlations analysis of final variables

	RNT_MEDI	ROAD_DEN	R_PUBWA	R_FARMI	R_BOTGA	R_RNT	POP_DEN
RNT_MEDI	1.000	0.104	0.079	−0.230	−0.277	−0.416	0.020
ROAD_DEN	_	1.000	0.104	−0.427	−0.489	0.401	0.470
R_PUBWA	_	_	1.000	−0.154	−0.094	0.040	0.024
R_FARMI	_	_	_	1.000	0.501	−0.164	−0.220
R_BOTGA	_	_	_	_	1.000	−0.156	−0.237
R_RNT	_	_	_	_	_	1.000	0.457
POP_DEN	_	_	_	_	_	_	1.000

On the basis of the census tract, we did cluster analysis by running K-Means in SPSS statistics software. We had near 2000 points in our dataset, but it was too large to do Hierarchical Cluster Analysis for judging the reasonable clusters existing in the dataset, so we run K-Means to assess the reasonable hierarchical clusters from 2 to 9 (it became obviously unreasonable above 6 clusters). We found no matter how many clusters we divided, there were two points existing in a unique cluster, the two points were testified as exceptional points, and eliminated because their households were very small (3 and 5), but with so large population (1085 and 1362), the public water, gas consuming per household all were beyond the normal values that formed a unique class. It

was obviously a problem caused by data quality.

We run K-means again from the 1924 tracts after the two points were eliminated. It also became definitely unreasonable above 6 clusters if tested from 2 to 9 clusters. The good result appeared as we run K-means in 4 clusters (Fig. 2).

4 RESULTS AND DISCUSSION

4.1 Results

From the scheme of 4 clusters, the 76 tracts (pink color) in the cluster 1, obviously as the urban center, were located only in central San Francisco and East Bay;

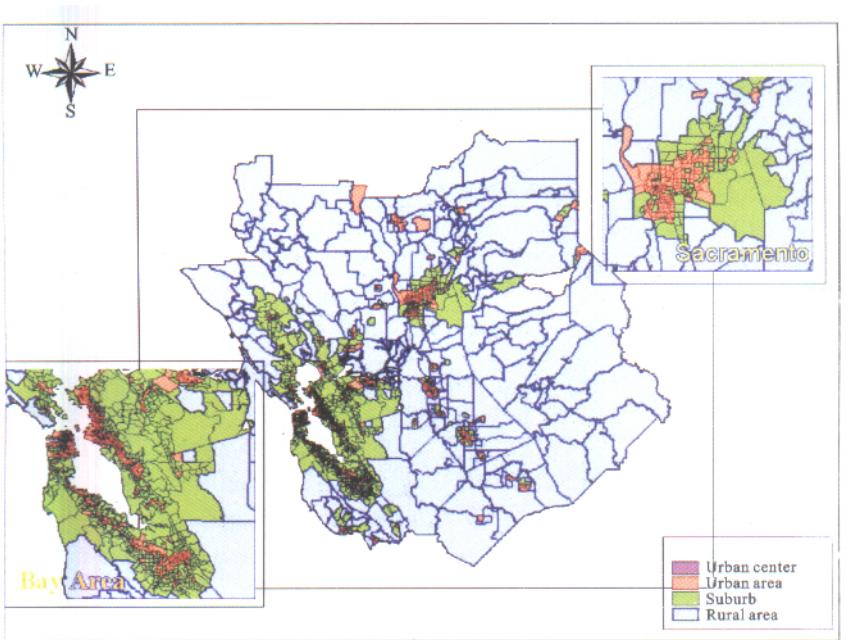

Fig. 2 Initial classification map

the 831 tracts (red color) as urban area in the cluster 2 were located mainly around Bay Area, Sacramento, and some scarcely locations in middle west and north; the 756 tracts (green color) in cluster 4, obviously as the suburb, were located mostly in the out edge of the cluster 2; it was reasonable that the suburban as edge of the urban was obviously enlarged in the metro-city such as Bay Area and Sacramento; and the 232 tracts (blue color) in the cluster 3, as rural areas, were located in most area of study area, beyond the suburban polygons (Fig. 3).

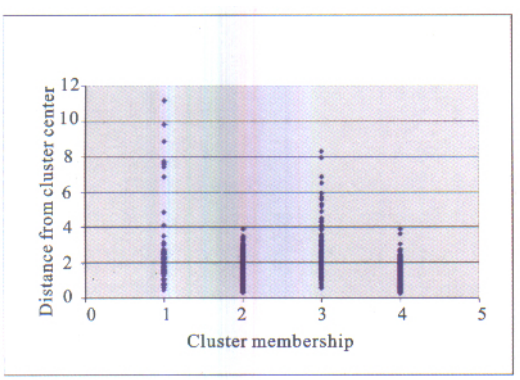

Fig. 3 Cluster result analysis

Most adjacent tracts were clustered very well, especially in the category of the urban and suburban. Only 29 tracts missed from total 1924 tracts in the study area; among 1895 valid tracts, only 11 tracts had large bias (above 6 from their cluster center); 1873 (99 %) tracts with the low value (below 4) biasing from the cluster center (Table 4 and Fig. 3). On the view of distance between final cluster centers, just cluster 1 biased little more from the others, and the distances between all other cluster centers were nearly equal (Table 5). So we obviously got a good clustering result.

Table 4 Number of cases in each cluster

Cluster	Cases
1	76
2	831
3	232
4	756
Valid	1895
Missing	29

Table 5 Distances between final cluster centers

Cluster	1	2	3	4
1	–	5.133	7.770	6.316
2	5.133	–	3.948	2.052
3	7.770	3.948	–	3.605
4	6.316	2.052	3.605	–

4.2 Discussion

Although most adjacent tracts were clustering very well, there were some unique tracts isolated in the other category. This could be explained as the phenomena of vacant lands in the urban or suburban area such as parks, animal zoos or harbors, etc., which would be shown up as the single tract with rural distribution in the urban or suburban category; on the contrary, the phenomena of a small town in the rural area could be shown up as the single urban tract surrounded by the rural category(Fig. 4). So we need to adjust our initial classification scheme produced by cluster analysis to be more reasonable. The method for such adjusting was to change category status of some isolated tracts (DAVID, 1998; BIN *et al*., 2000). The principle for adjustment was: the single tract would be merged into surrounding, except for the non-single isolated tract(above two in adjacency). Among the 1924 tracts, only small parts of tracts (total 76 tracts) need to change their category status, and then we got the final map of classification (Fig. 5).

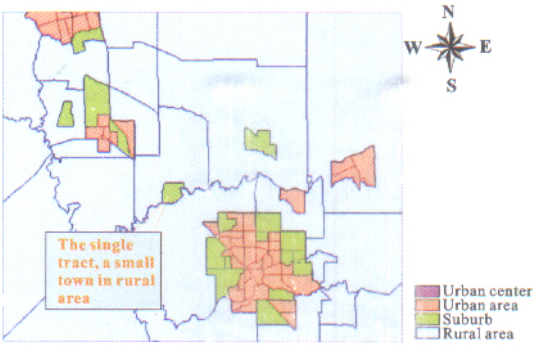

Fig. 4 A small town in rural area

The final result was compared with classification scheme of TIGER (Fig. 6). We observed most urban polygon matched with the classification result very well except for some small polygons in the north and middle. It was reasonable that the urban boundaries of our result mostly smaller than the TIGER′s, since no suburban class existed in the TIGER's scheme.

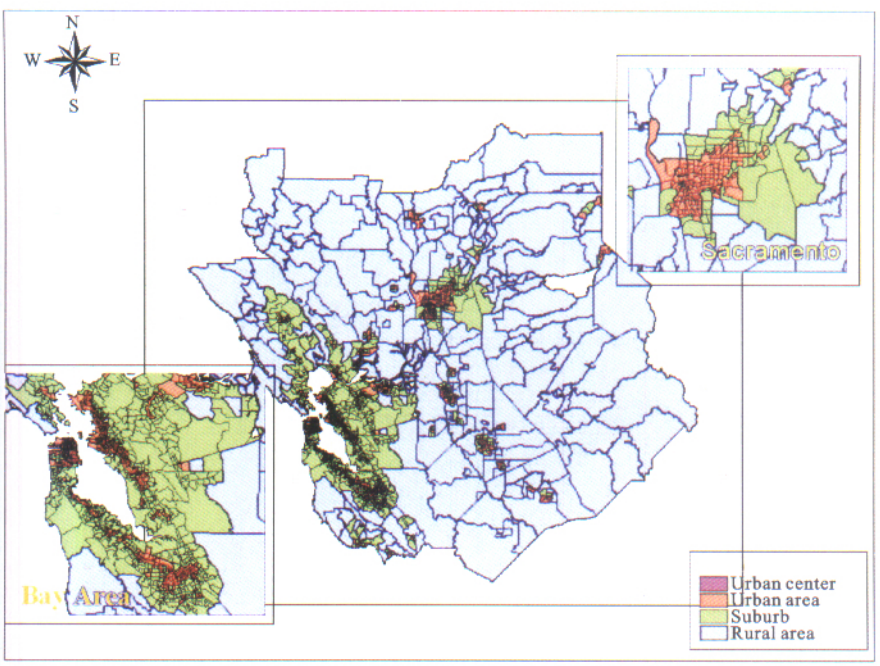

Fig. 5 Final classification map

5 CONCLUSIONS

Based on the spatial characteristics of 7 multi-variables (population density, ratio of public water consuming, road density, ratio of bottle gas consuming, ratio of renters, ratio of farm income, rent price), 1924 tracts, which composed 28 counties in the central California, were clustered into 4 classes so that we could get the urban, suburban and rural category in the region. The urban area were mostly located in the Bay Area and Sacramento; the suburban area were located mostly at the outer edge of urban area, and in the metropolitan area, the suburb extends larger; the rural area occupies most other locations in the study area.

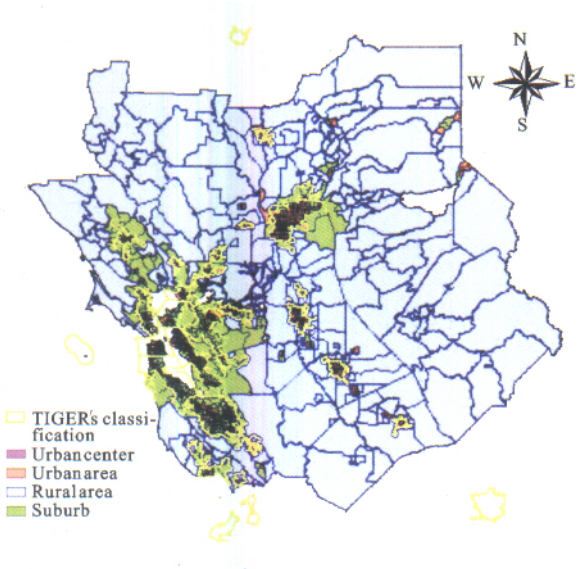

Fig. 6 Comparison between our classification scheme and TIGER′s

ACKNOWLEDGEMENTS

I am grateful to Prof. John Radke for providing me with consistent encouragements and valuable suggestions. Prof. PU Rui-liang and Prof. YANG Kai-zhong (Peking University, P. R. China) gave some good suggestions. This paper was finished during the first author′s academic visiting in the GISC, UC, at Berkeley.

REFERENCES

ALAN T, VLADIMIR E, 1998. Cluster discovery techniques for exploratory spatial data analysis [J]. *Geographical Information Science*, 12(5): 431–443.

ANDRE S, 2000. Land readjustment and metropolitan growth: an examination of suburban land development and urban sprawl in the Tokyo metropolitan area [J]. *Progress in Planning*, 53 (4): 217–330.

BIN J, CHRISTOPHE C, BJORN K, 2000. Integration of space syntax into GIS for modeling urban spaces [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2(3–4): 161–171.

DAVID M, 1998. Automatic neighborhood identification from population surfaces [J]. *Computers, Environment and Urban Systems*, 22(2): 107–120.

HATHOUT S, 2002. The use of GIS for monitoring and predicting urban growth in East and West St Paul, Winnipeg, Manitoba, Canada [J]. *Journal of Environmental Management*, 66(3): 229–238.

LAWRENCE D, 2003. Community segmentation: a new system for geodemographic analysis [J]. *Arc News*, 25(3): 1–2.

MARET I, DAKAN B, 2003. GIS and sprawl management [J]. *Bulletin-d′-Association-de-Geographes-Francais*, (2): 220–234.

NELSON A C, 1992. Characterizing exurbia [J]. *Journal of Planning Literature*, 6(4): 350–368.

U.S.A. Census, 1995. Census Tiger 1995 Data [EB/OL]. http://www.gisc.berkeley.edu/data/tiger1995/U.S.A. Census, 2000.

U.S.A. Census, 2000. Geographic Terms and Concepts [EB/OL]. http://www.census.gov/geo/www/tiger/glossry2.pdf

WANG Qiao, YAN Shou-hu, ZHAO Jian, 1998. The model of clustering for classification [A]. In: *Modeling and Management of Geographic Information Science Applied by Regional Planning* [C]. Beijing: Yuhang Press, 55–57. (in Chinese)

ZORICA N B, MARY E F, ABBAS R *et al*., 2004. Are SDIs serving the needs of local planning? Case study of Victoria, Australia and Illinois, USA [J]. *Computers, Environment and Urban Systems*, 28(4): 329–351.