HHS Public Access

Author manuscript

Int Conf Geoinform. Author manuscript; available in PMC 2018 January 31.

Published in final edited form as:

Int Conf Geoinform. 2015 June ; 2015: .

Understanding the Clustering Patterns in Physician Distribution Through Affinity Propagation

Xuan Shi[1,*], **Bowei Xue**[1], and **Imam Xierali**[2]

[1]Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, U.S.A

[2]Association of American Medical Colleges, 655 K Street NW, Ste. 100, Washington, DC 20001, U.S.A

Abstract

The spatial distribution of physicians has a significant impact in public health research. It is critical to clarify whether the addresses provided by the physicians are the home addresses or the practice addresses, since the practice address is the key to understand relevant issues of maldistribution, accessibility and disparity. Through a pilot study as partial effort of the research project “Reducing Physician Distribution Uncertainty in Spatial Accessibility Research” sponsored by the National Institutes of Health (NIH award number 1R21CA182874-01), appropriate solutions were developed to differentiate the home addresses from practice addresses. This paper introduces how to understand the clustering patterns in physician distribution through Affinity Propagation, a relatively new clustering algorithm, to derive the potential extent of the practice locations for those physicians who provided home addresses. The physician data is derived from the 2014 American Medical Association (AMA) Physician Masterfile, while two counties (Fulton and DeKalb) in the metropolitan area of Atlanta, Georgia were selected as the study area. Both Euclidian distance and driving distance were applied in the AP algorithm, while gravity models based AP calculation were applied in comparison to the clustering of individual physicians. By justifying preference and similarity parameters in the AP calculation, hierarchical clustering patterns can be derived and perceived. Future research challenges in AP clustering are identified, while this pilot study can be extended with broader impact in public health research.

Keywords

spatial cluster analytics; Affinity Propagation; physician distribution

I. Introduction

Physician shortage due to the maldistribution of physician workforce is a significant feature and concern of the US health care system [1–2]. Consequently health disparity issues have been argued particularly when the shortage of physicians would be obvious and serious in certain areas [3–5], while federal programs, such as Health Professional Shortage Areas (HPSAs) and Medically Underserved Areas and Populations (MUAPs), were established to

*Corresponding author, xuanshi@uark.edu.

handle the spatial maldistribution of health care resources [6]. For the general public, access to medical or health service providers has been critical to the patients, especially given proposed coverage expansion in the Patient Protection and Affordable Care Act extended to millions of Americans.

The spatial distribution of physicians can be perceived and analyzed from a variety of data sources, such as the Physician Masterfile generated by the American Medical Association (AMA), the National Provider Identifier (NPI) database, and Georgia State Licensure data. Such data sources contain the information about the addresses of the physicians who responded to the surveys conducted by different professional associations and administration agencies.

Besides the traditional inaccuracy problems in the address information provided in the survey or the spatial errors in the geocoding results [7–12], however, it is of particular significance and interest to clarify whether the physician's address is the home address or the practice address for clinics or relevant healthcare services. In our pilot research, we successfully differentiate the home addresses from non-home addresses through spatial analytics, text mining, and visual examination [13] by parsing and comparing the physicians' addresses to the addresses documented in the parcel datasets that contain the zoning code of the land use types. When the physician's address can be matched to a parcel, it can be clarified whether it is home address or not as the zoning code of the parcel specifies its usage as single family or multi-family resident community, for commercial or industrial development, or as tax exempted areas including governmental properties, churches, hospitals, non-profit organizations, and so on.

This paper extends the prior research results and introduces how to estimate the potential extent of the practice locations for those physicians who provided home address by analyzing the clustering patterns in physician distribution through Affinity Propagation (AP). Initial outcome and major research challenges are discussed in this paper to promote optimized solution.

II. Study area, data and prior results

As a pilot research project, two counties, Fulton and DeKalb, in the metropolitan area of Atlanta, Georgia were selected as the study area since both counties have parcel data available for this research. In the year of 2014, Fulton County has 353,723 parcel polygons, while DeKalb County released 235,681 parcels as point features.

Considering the annual AMA Masterfile contains the most comprehensive information about physicians in the U.S., the physician data used in this research is extracted from the 2014 AMA Masterfile, which includes information about the physicians' addresses, practice type, specialty, age, gender, employment settings, primary professions, medical school, Graduate Medical Education (GME) ending date, residency institutions, and other data. A total of 6,271 physicians located within Fulton and DeKalb counties are further identified and extracted.

Along with many other attribute items, the parcel datasets contain the information about the address and land use type of the parcel. The zoning code for the land use class of the parcel could reveal whether the parcels are used as agricultural, business, commercial, industrial, residential, exempted areas, or for other purposes. When the physician's address is matched to a parcel, the class code in the parcel datasets helps to differentiate whether the physician's address is a home address or not by checking whether the corresponding parcel is classified as resident area or not.

The result of address matching is summarized in Table 1. Among 6,271 physicians in Fulton County and DeKalb County, a small amount of physicians (about 8%) did not provide any address information or their addresses are not valid. Within the remaining 92% of physicians, about 4% of the addresses cannot find corresponding counterparts in the parcel datasets. While 88% of physicians' addresses could be matched to parcel addresses, 81% can be identified as practice addresses, 6% as home addresses, and 1% could not be determined when multiple parcels share the same street address but have different land use codes. Within the total number of matched addresses, 30% of the home addresses (i.e. 121) have an exact match with the addresses in the parcel data, 57% of the practice addresses (i.e. 2,907) have an exact match with the addresses in the parcel data.

III. Cluster analysis by affinity propagation

One of the aims of the R21 research project was to model or estimate the physician practice site selection and assign mislocated physicians to clinic clusters with the assumption that physicians with unknown practice locations would work in an existing health care cluster in the study area. It will also be assumed that the probability of a physician working at a certain cluster site is determined proportionally by the attractive characteristics of the site/cluster and inversely by the travel impedance between the physician location and the site. The denser clusters will generally attract more subspecialists than primary care physicians, while the looser clusters will typically contain more primary care physicians. Similarly, it is assumed that more specialized physicians would choose to work in larger and more complex health clusters.

Among varieties of classification and clustering approaches for spatial data mining and knowledge discovery [14], this research applied the Affinity Propagation (AP) [15] approach for several reasons. Unlike other classification or clustering algorithms, such as ISODATA, k-means, and Maximum Likelihood Classifier, AP does not specify a pre-defined arbitrary number of clusters in advance but will derive the number of clusters as the result. In comparison to other raster-based clustering approaches, such as Hot-spot analysis, AP can generate a clear and concrete connection for each feature to its given cluster center. For this reason, AP has significant potential in the identification of spatial clusters and other research and applications, such as data resampling, spatial filter, and pattern analysis. Particularly in this research, AP helps to estimate the physician practice site selection and assign mislocated physicians to clinic clusters.

The AP algorithm was discussed in [15]. A similarity matrix S contains n x (n−1) records of the negative values of the distance between each point to all other points. The other input

data contains the preference value of the n input points. The similarity matrix S describes how each data point is presented to be the exemplar, while data points with higher preference values could be selected as cluster centers or exemplars. In this case, the preference value determines the number of identified clusters. In AP, all data points are considered equally as potential exemplars or the cluster centers. For this reason, the preference values are initialized to a common value, which is usually the median in the similarity matrix. In general, AP is an optimization process to maximize the similarity or to minimize the total sum of intra-cluster similarities.

IV. Cluster patterns by different approaches

For all of 6,271 physicians in the two counties of the study area, only 1,056 the unique locations of the physicians can be derived by comparing and examining the x and y coordinates. For this reason, each point in Figure 1 actually is not equivalent. Figure 2 would help to understand the concentration scale of the physician distribution in the study area. A large cluster of physicians normally means a critical medical center in comparison to the small clinic locations. However, the hierarchical clusters derived from AP may display different patterns.

In the generic AP calculation, similarity was measured by the distance between the features. By justifying preference and similarity parameters in the AP calculation, hierarchical clustering patterns can be derived and perceived based on individual physician locations. Table 2 contains 1 – 9 physician clusters derived from AP by applying Euclidian distance in the calculation, while Table 3 contains 1 – 9 physician clusters derived from AP by applying driving distance in the calculation. When driving distance is applied, the clustering patterns are slightly different from those based on Euclidian distance. In comparison to the concentration of physicians at different locations displayed in Figure 2, the center of clusters derived from AP are not necessarily at the locations of clinic or medical centers that would host more physicians.

A simple gravity model is constructed based on the number of physicians at each of those 1,056 the unique locations. AP is calculated again using Euclidian distance and driving distance. The results are displayed in Table 4 and 5. The clustering patterns are significantly different from those displayed in Table 2 and 3 correspondingly. When more clusters are derived, the locations of the parent cluster centers seem to be preserved. Particularly the center of the clusters derived from AP seems to be more meaningful in connection to the distribution of physicians displayed in Figure 3.

V. ESITMATION OF THE POTENTIAL PRACTICE LOCATION

Now that physician clusters derived from AP by applying driving distance over gravity model could be more reasonable to reflect the distribution and clustering of physicians, such an approach could be used to estimate the potential practice location for those physicians who provided home addresses. Figure 3 displays 15 physician clusters derived from AP by applying driving distance over gravity model. Those little red triangles represent the location of physicians who provided home addresses. The spatial extent of their potential practice

location could be determined by the cluster it belongs to. If the physician is not located within the corresponding cluster, he or she may join a nearby cluster. Such a visualized approach will contribute to fulfill one aim of this research in which the Huff model will be applied based on the hypothesis or probability that a mislocated physician with a set of alternative practice sites will select a particular site is directly proportional to the perceived utility of each alternative site [16].

VI. Conclusions

Affinity Propagation seems to be an amazing approach to understand the clustering patterns in physician distribution. The derived pattern helps to estimate the spatial extent of their potential practice locations when some physicians provided home address, while it helps to validate the result to be determined by the Huff model. Knowledge gained via this research will help address the maldistribution of health care providers by assuring better health workforce distribution information and by more accurately measuring spatial accessibility to physician services. The study method is scalable from the local level to regional and national levels. It can also be replicated in many other types of health workforce data such as the location of nurse practitioners and physician assistants. Decision-making for assessing physician accessibility will thus become more dynamic and comprehensive in response to different patterns of location allocation with multiple possibilities.

Although AP has obvious advantages in comparison to many other approaches for clustering analysis, it was noticed that AP's computational and memory requirements scale linearly with the number of similarities input; for non-sparse problems where all possible similarities are computed, these requirements scale quadratically with the number of data points. Exploring scalable solutions for AP calculation over big data will make AP more powerful and capable to handle real world problems in the future.

Acknowledgments

This research was supported by the National Institutes of Health through the award NIH 1R21CA182874-01.

References

1. Council on Graduate Medical Education (COGME). Physician Distribution and Health Care Challenges in Rural and Inner-City Areas, 10th Report. Rockville, Md: COGME; 1998.
2. Council on Graduate Medical Education (COGME). Advancing Primary Care, 20th Report. Rockville, Md: COGME; 2010.
3. Ricketts TC, Goldsmith LJ, Randolph R, Lee R, Taylor DH, Ostermann J. Designating places and populations as medically underserved: A proposal for a new approach. Journal of Health Care for the Poor and Underserved. 2007; 18(3):567–589. [PubMed: 17675714]
4. Zhang X, Phillips RL, Bazemore AW, Dodoo MS, Petterson SM, Xierali I, Green LA. Physician distribution and access: Workforce priorities. American Family Physician. 2008; 77(6):1378. [PubMed: 18853533]
5. Peterson LE, Bazemore AW, Bragg EJ, Xierali IM, Warshaw G. Rural urban distribution of the US geriatrics physician workforce. Journal of the American Geriatrics Society. 2011; 59:699–703. [PubMed: 21438865]

6. Xierali I, Bazemore AW, Phillips RL, Petterson SM, Dodoo MS, Teevan B. A perfect storm: changes impacting Medicare threaten primary care access in underserved areas. American Family Physician. 2008; 77(12):1738. [PubMed: 18853534]
7. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003; 14(4):408–412. [PubMed: 12843763]
8. Bichler G, Balchak S. Address matching bias: ignorance is not bliss. Policing: An International Journal of Police Strategies & Management. 2007; 30(1):32–60.
9. Drummond WJ. Address Matching: GIS Technology for Mapping Human Activity Patterns. Journal of the American Planning Association. 1995; 61(2):240–251. http://doi.org/10.1080/01944369508975636.
10. Goldberg DW, Wilson JP, Knoblock CA. From text to geographic coordinates: the current state of geocoding. URISA Journal. 2007; 19(1):33–46.
11. Goldberg, DW., Jacquez, GM., Mullan, N. Geocoding and Health. In: Boscoe, F., editor. Geographic Health Data: Fundamental Techniques for Analysis. CABI Press; Wallingford, UK: 2013.
12. McLafferty S, Freeman VL, Barrett RE, Luo L, Shockley A. Spatial error in geocoding physician location data from the AMA Physician Masterfile: implications for spatial accessibility analysis. Spatial and Spatio-Temporal Epidemiology. 2012; 3(1):31–38. [PubMed: 22469489]
13. Shi X, Xue B, Xierali I. Reducing the Uncertainty in Physician Distribution through Spatial Analytics, Text Mining, and Visual Examination. 2015 Under review.
14. Guo D, Mennis J. Spatial data mining and geographic knowledge discovery – An introduction. Computers, Environment and Urban Systems. 2009 Nov; 33(6):403–408.
15. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007 Feb. 315:972–976. [PubMed: 17218491]
16. Huff DL. A Probabilistic Analysis of Shopping Center Trade Areas. Land Economics. 1963; 39:81–90.

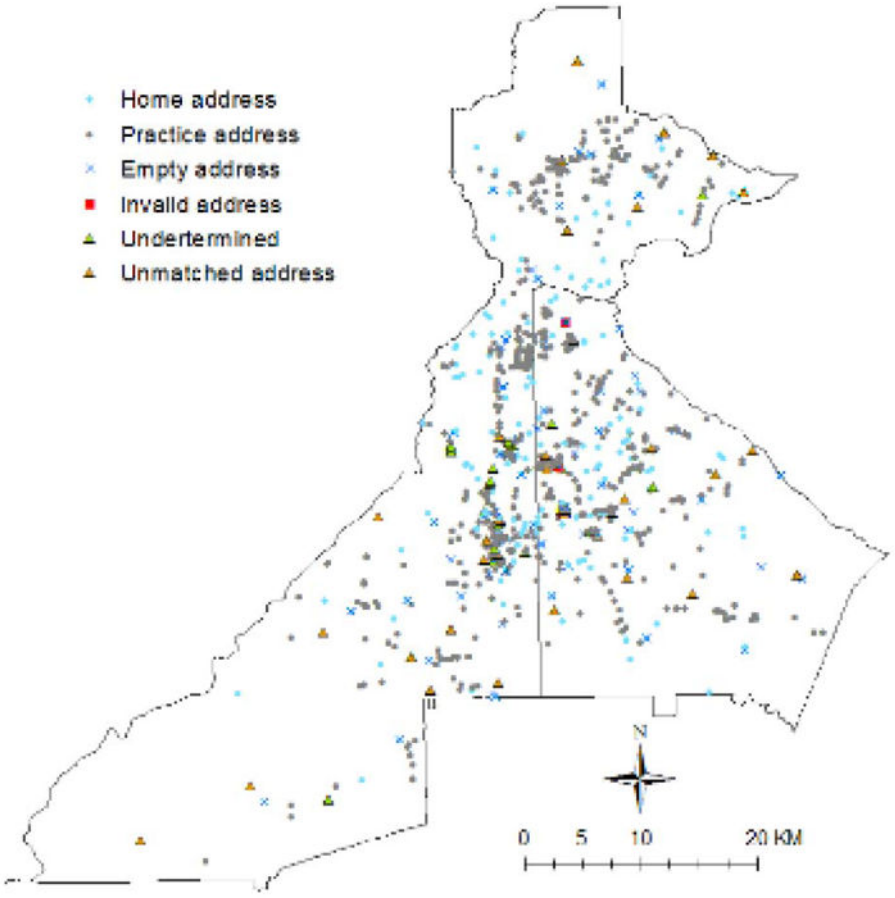

Figure 1.
Physician locations

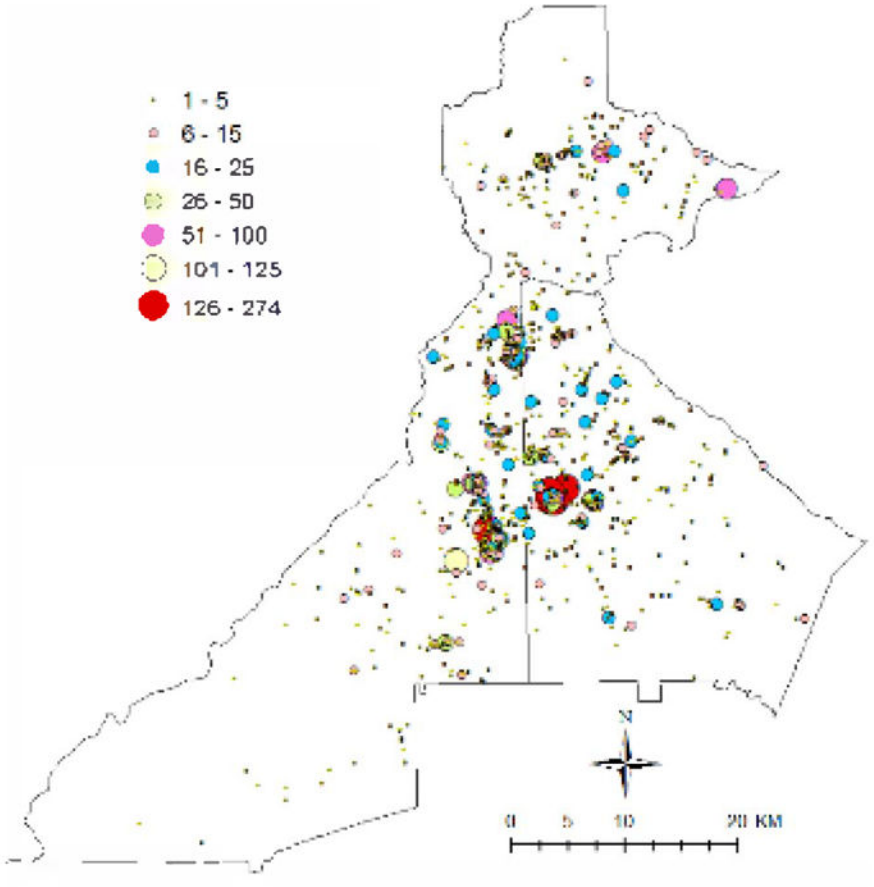

Figure 2.
Concentration of physicians

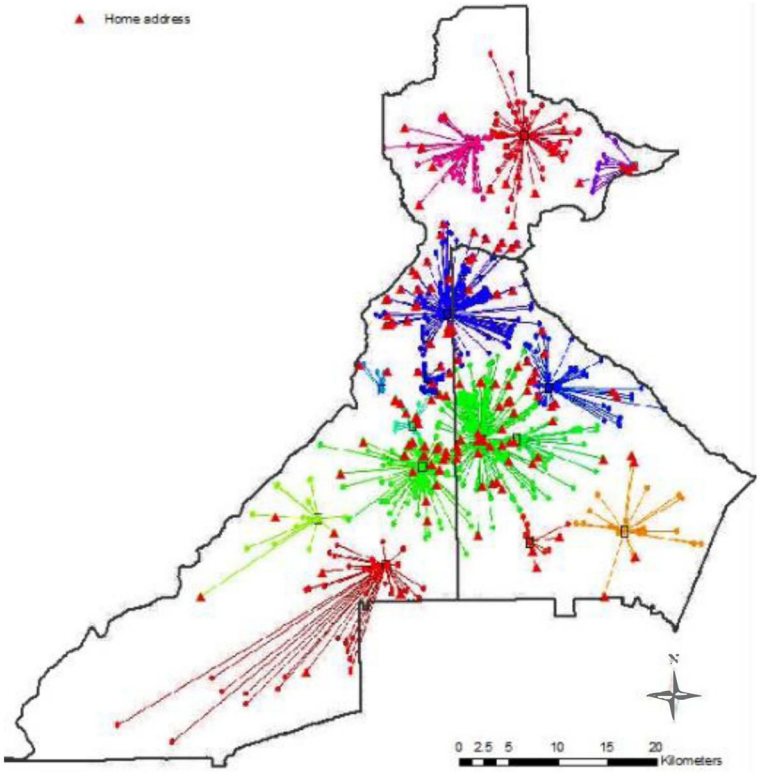

Figure 3.
15 physician clusters derived from AP by applying driving distance over gravity model

TABLE I

Address matching results

Empty addresses		488	7.78%
Invalid addresses		18	0.29%
Matched addresses	Residential	401 (121)	6.39%
	Non-residential	5,063 (2,907)	80.74%
	Undetermined	50	0.80%
Unmatched addresses		251	4.00%
Total		6,271	100%

TABLE II

Physician clusters derived from AP by applying Euclidian distance

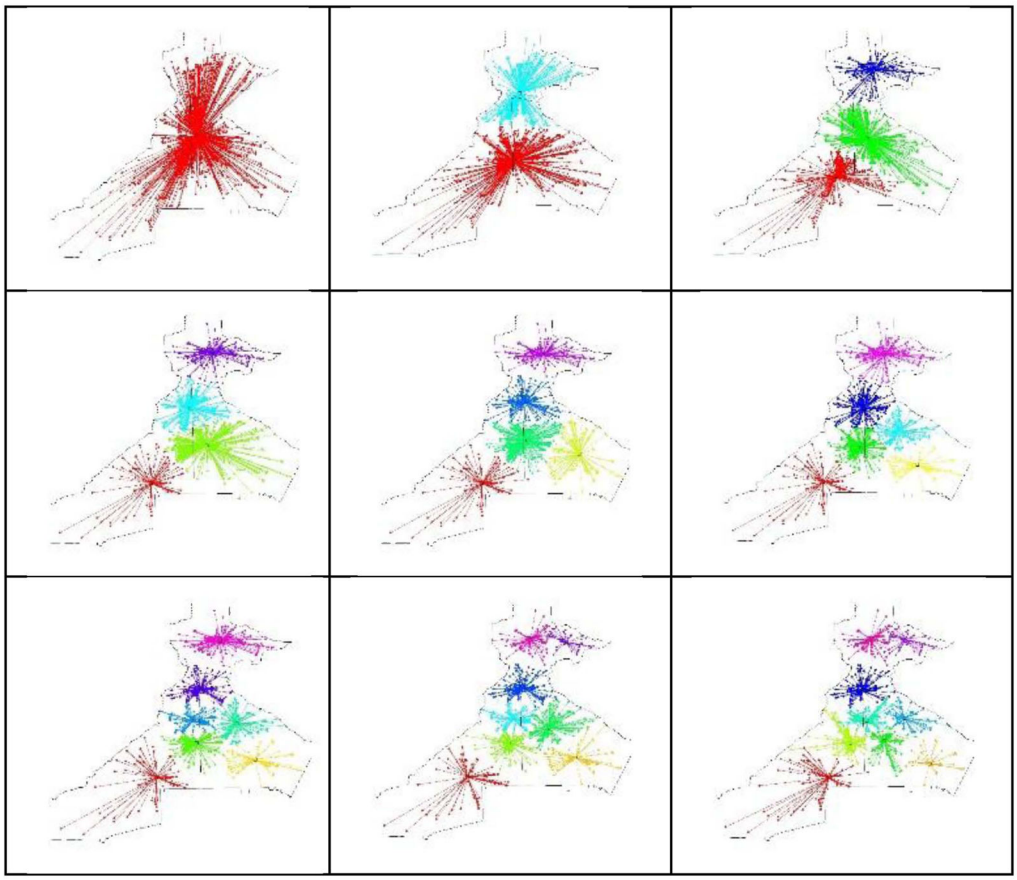

TABLE III

Physician clusters derived from AP by applying driving distance

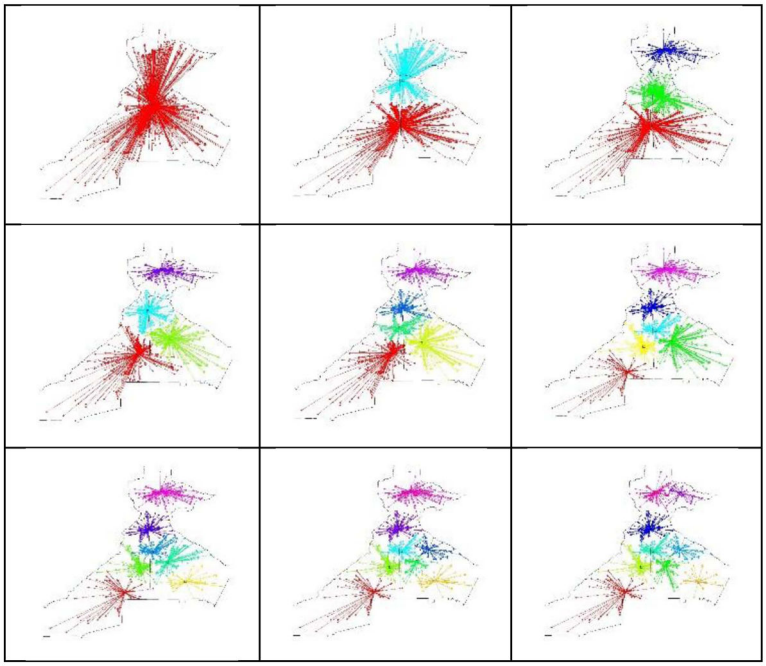

TABLE IV

Physician clusters derived from AP by applying Euclidian distance over gravity model

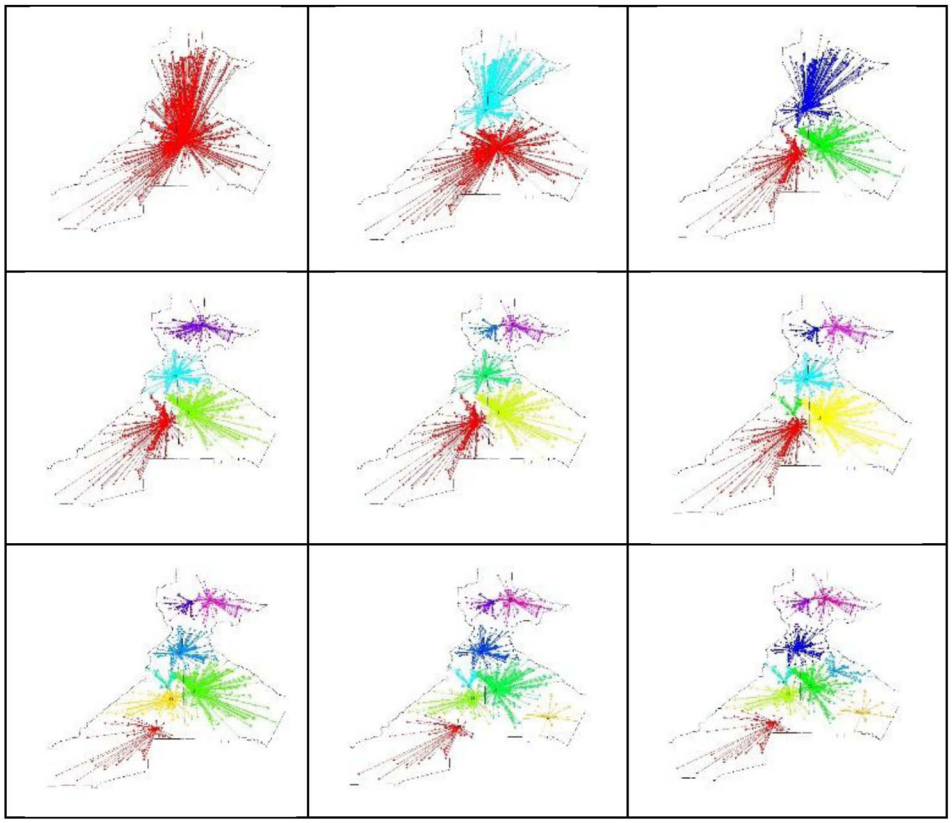

TABLE V

Physician clusters derived from AP by applying driving distance over gravity model