Analyzing the correlations between the uninsured and diabetes prevalence rates in geographic regions in the United States

Xiao Luo
Purdue School of Engineering and Technology, IUPUI
Indianapolis, Indiana 46202-5160
Email: luo25@iupui.edu

***Abstract*—The increasing prevalence of diagnosed diabetes has drawn attentions of researchers in recently years. Research has been done in finding the correlations between diabetes prevalence with socioeconomic factors, obesity, social behaviors and so on. Since 2010, diabetes preventive services have been covered under health insurance plans in order to reduce diabetes burden and control the increasing of diabetes prevalence. In this study, a hierarchical clustering model is proposed by using Expectation-Maximization algorithm to investigate the correlations between the uninsured and diabetes prevalence rates in 3142 counties in United States for years from 2009 to 2013. The results identified geographic disparities in the uninsured and diabetes prevalence rates of individual years and over consecutive years.**

I. Introduction

Researches and statistics have shown that diabetes prevalence rate has been increased significantly in the world over the last decade [1][2][3], and will continue to increase rapidly over years to 300 million cases in 2025 [4]. Some research have been done to investigate associations between health insurance and patients with diabetes [5][6]. These research mainly concentrated on evaluating health insurance on providing treatment and access to the care for the patients if they have been diagnosed with diabetes. The research results provided valuable information into the effectiveness of health insurance in the reactive care model of health care. Starting from late 2010, certain preventive health care services are covered under some health insurance [10][12]. Diabetes screening is the one of the preventive services that are under the coverage. However, this study is the first to investigate the correlation between health insurance (uninsured rate) and the diabetes prevalence in geographic regions by using machine learning algorithm. The previous related research used statistic analysis, such as regression analysis [11] to determine the correlation between diabetes prevalence with other socioeconomic factors. The results of this research will shed the light on whether the improving insurance rate (reducing uninsured rate) has a positive correlation with reducing diabetes prevalence rate, especially when diabetes screening is under the coverage of health insurance starting from 2010.

In this research, we proposed a hierarchical clustering model based on Expectation-Maximization (EM) algorithm to identify the geo-clusters on the U.S. map based on the correlations between diabetes prevalence rate and uninsured rate. The experimental data sets included in this research are the estimates of the prevalence of diagnosed diabetes that was published by Centers for Disease Control and Prevention (CDC) [13] and the estimates of the uninsured that was published by United States Census Bureau [14]. The first level EM model was employed to identify the county-level geo-clusters based on the correlations between diabetes prevalence and uninsured rate of an individual year. The second level EM model is based on the results of the first level, which is used to identify the geographic regions that have similar correlations between these rates over consecutive years. The results show that there are disparities in the correlations in different geographic regions in each year. The correlation analysis over the years shows that counties in the North Western and Western regions have some negative correlations between the two rates, whereas counties that are in the South and South Eastern regions have some positive correlations between the two rates.

The rest of paper is organized as follows. The related work in the literature is summarized in Section II. The hierarchical Expectation-Maximization clustering model and Expectation-Maximization algorithm are detailed in Section III. Section IV presents the experimental results and analysis. Finally, conclusions are drawn and the future work is given in Section V.

II. Background and Related Work

Researches have shown that disparities in the diabetes prevalence in geographic regions associated with many factors, such as socioeconomic factors [7][8], ethnicity [3][9] and so on.

In 2011, Barker et al. [15] published a paper that identified the diabetes belt on the U.S. map using the Behavioral Risk Factor Surveillance System (BRFSS) data of 2007 and 2008. The diabetes belt was located in the southeast region. They also concluded that the people in the diabetes belt were more likely to be non-Hispanic African-American and were those who lead an inactive life style and be obese. This research analyzed the diabetes prevalence in association with factors, such as education, gender, life style, ethnicity and so on.

Zhang et al. [5] analyzed the relationships between access to the health care insurance and control of A1C, blood pressure

This is the author's manuscript of the article published in final edited form as:
Luo, X. (2017). Analyzing the Correlations between the Uninsured and Diabetes Prevalence Rates in Geographic Regions in the United States. In 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 44–50). https://doi.org/10.1109/CHASE.2017.58

and lipid of the population who have developed diabetes by using statistic analysis. They found that 84% of adults are insured and 16% of adults are uninsured in U.S. The uninsured population more likely to have fair or poor health.

Mokdad et al. [9] analyzed changes in diabetes prevalence from 1990 to 1998 by selected characteristics and state. The characteristics included age-group, sex, education level, and so on. They have found that of the 43 states in the comparison, 35 showed an increase in the diabetes prevalence.

Pickle and Su [8] showed the correlations between health insurance coverage and health risk factors such as smoking, obesity, and mammography for year 1992 to 1998 by county. Geographic patterns had been shown based on each individual factor. Diabetes prevalence factor were not discussed.

Michimi and Wimberly [16] studied the geographic patterns of obesity and associated risk factors include physical activity, fruit and vegetable consumption in the U.S.. The Behavioral Risk Factor Surveillance System (BRFSS) data from 2000 to 2006 was used for the analysis. Although the diabetes prevalence rate was not directly involved, the obesity is one of the major risk factors for type 2 diabetes prevalence. Their results showed that higher prevalence of obesity was found in the counties of the South, whereas lower prevalence was found in the West and the Northeast.

To the best of the author's knowledge, no research has done in investigating the correlation between the uninsured and diabetes prevalence for geographic regions by using machine learning algorithms. This is the first research to provide synoptic picture of the correlation and the spatial patterns in U.S..

III. Methodology

The main objective of this research is to exploring whether there are the spatial patterns of correlation between diabetes prevalence and the uninsured before and/or after diabetes preventive service was covered under insurance. To this end, a hierarchical clustering model is proposed by using the Expectation-Maximization (EM) algorithm. The first level EM clustering model is to identify the geo-clusters based on the diabetes prevalence rate and uninsured rate for an individual year. The second level EM clustering model is built upon the first level EM clustering model to identify geo-clusters based on the correlations over consecutive years. Figure 1 demonstrates the proposed system framework. The description of the hierarchical clustering model and the Expectation-Maximization (EM) algorithm is given in the following sub-sections.

A. Overview of the proposed hierarchical clustering model

The inputs to the first level EM clustering model are two dimensional vectors ($D = \{v_1, v_2\}$). One dimension (v_1) presents the diagnosed diabetes prevalence rates (DPRs), whereas the other dimension (v_2) presents the uninsured rates (URs) of the corresponding counties. The outputs of first level clustering modeling include a number of clusters with their cluster centers. The number of clusters and the cluster distributions are different from one year to another. In figure 1, the syntactic data is used to demonstrate the cluster distributions of each year. The correlation coefficient is used to calculate the relationships between the DPRs and URs within clusters. The Eq. 1 demonstrates the calculation of the correlation coefficient ($Correl(X, Y)$) for two variables: $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, where $Cov(X, Y)$ is the covariance of the variables X and Y, and S_X and S_Y are standard deviations of X and Y respectively.

$$Correl(X, Y) = \frac{Cov(X, Y)}{S_X S_Y} \tag{1}$$

Although the results of first level clustering provide the geo-clusters and correlations between DPRs and URs within clusters for every year. The correlations over consecutive years are not provided. To this end, the second level EM clustering model is proposed. For a specific year, the cluster center ($\{v_{c1}, v_{c2}\}$) of the cluster (C) that an input vector belongs to is used to present the input. The Piecewise Linear Approximation (PLA) [23] slope ($PLA.slope$) is employed to calculate the correlations between DPRs and URs of two consecutive years from t to $t + 1$. If the $PLA.slope$ is below 0, DPR and UR are in positive correlation. Otherwise, DPR and UR are in negative correlation. The PLA slope metric ($PLA.slope$) is given as Eq. 2. The D^t and D^{t+1} are the input vectors that are consisted of DPRs and URs for a county in year t and $t + 1$. The first level clustering results of year t and $t + 1$ identify that D^t belongs to cluster C^t and D^{t+1} belongs to C^{t+1}. The vectors $\{v_{c1}^t, v_{c2}^t\}$ and $\{v_{c1}^{t+1}, v_{c2}^{t+1}\}$ are the values of cluster centers of C^t and C^{t+1} respectively.

$$PLA.slope = \frac{v_{c1}^{t+1} - v_{c1}^t}{v_{c2}^{t+1} - v_{c2}^t} \tag{2}$$

A vector is constructed to present the correlations in more than two consecutive years. Each unit in the vector is to present $PLA.slope$ values between two consecutive years. The syntactic data in figure 1 shows that given three clustering results based on data of three years, there will be a vector of dimension of two to present the correlations over the years. Given syntactic data of five counties, based on the vector of the $PLA.slope$ values, they could be clustered into three clusters (Green, Red and black).

B. EM clustering algorithm

The Expectation-Maximization (EM) clustering algorithm [18] has been widely used in the health informatics domains for pattern recognition [19] [20], especially when the number of clusters is unknown. There are two major steps for the EM algorithm [18]: E-step and M-step. The E-step is to compute the posterior distribution for all data points $\{x_1, x_2, \ldots, x_M\}$ by using the estimated hidden variables $\{C_1, C_2, \ldots, C_N\}$ and the parameter θ for the hidden variables. The posterior computing for a given data point x_i is given in Eq. 3

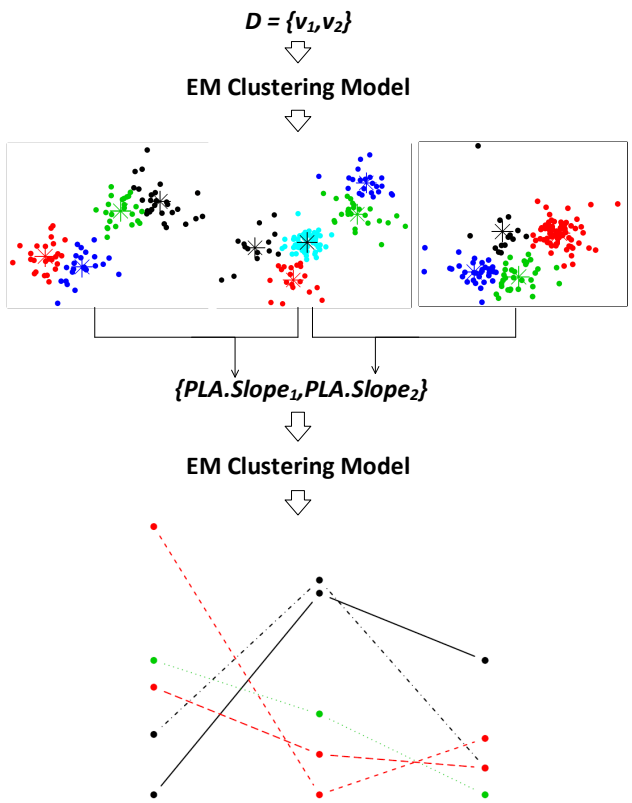

Fig. 1. Overview of the proposed hierarchical clustering model with demonstration of syntactic data

$$P(C_j|x_i,\theta) = \frac{P(x_i|C_j,\theta)P(C_j,\theta)}{\sum_{n=1}^{N} P(x_i|C_n,\theta)P(C_n,\theta)} \tag{3}$$

The M-step is using the posterior distribution that is calculated at in the E-step for the hidden variables, and optimize the parameter θ using the expected values of the hidden variables, as Eq. 4.

$$\theta' = \arg\max P(C|x,\theta) \tag{4}$$

The E-step and M-step iterate till there is no changes between the parameter θ and optimized θ' for two iterations or the number of iterations reaches the pre-defined maximum number of iterations. In this study, the EM algorithm implemented in Waikato Environment for Knowledge Analysis (WEKA) [22] was employed.

IV. Experimental Setups, Results and Discussion

In this research work, the published estimates of the prevalence of diagnosed diabetes by CDC [13] and the published Small Area Health Insurance Estimates (SAHIE) data [14] by United States Census Bureau have been used. The published estimates of the prevalence of diagnosed diabetes by CDC includes age-adjusted rate prevalence of diagnosed diabetes rates (DPR) from year 2004 to 2013. It concludes both Type 1 diabetes and Type 2 diabetes. However, based on CDC website [24], the Type 1 diabetes and Type 2 diabetes might account for about 5% and 90% of all diagnosed cases of diabetes respectively. So, this diabetes prevalence rate largely reflect the Type 2 diabetes prevalence. The SAHIE data contains uninsured rate (UR) of the population of age < 65 in U.S.. It provides detailed uninsured rates based on the counties, age, race, sex and income categories from year 2005 to 2014. In this research, the overall uninsured rates for counties without considering other factors are extracted. The population of age $>= 65$ are likely covered by medicare [5], and the uninsured rate basically reflect the uninsured rate of the whole population. In this research, data from 2009 to 2013 of both data sets are integrated. It covers the years before and after diabetes preventive services started to be covered under the health insurance in 2010. After removing counties that has no data of either diabetes prevalence rate or uninsured rate in any the selected years, there are total of 3142 counties included in this study. The commonwealths and territories, such as municipalities of Puerto Rico are not included in this study.

TABLE I
First level EM clustering result - 2009

	Clusters							
	0	1	2	3	4	5	6	Overall
# of Counties	780	155	504	617	279	129	678	3142
DPR								
mean	10.79	8.78	7.10	7.57	11.01	9.73	9.12	9.12
std. dev.	2.02	0.28	1.10	0.75	0.85	0.86	0.88	1.95
UR								
mean	20.66	28.95	19.24	11.56	18.38	25.59	15.38	18.27
std. dev.	1.40	5.39	3.81	2.46	2.07	6.25	1.88	5.71

TABLE II
First level EM clustering result - 2010

	Clusters									
	0	1	2	3	4	5	6	7	8	Overall
# of Counties	511	384	176	45	652	565	131	329	349	3142
DPR										
mean	8.70	7.49	6.52	11.23	11.44	10.24	9.94	9.01	7.37	9.33
std. dev.	0.89	0.63	1.04	2.65	1.37	0.96	1.27	0.47	0.73	2.01
UR										
mean	13.66	17.37	20.63	19.43	21.08	17.0	24.61	24.37	10.86	18.54
std. dev.	1.89	5.08	6.06	2.34	2.63	1.84	4.12	6.83	1.39	5.60

A. Results of the first level EM and Discussion

The first level EM clustering model has been trained on the data for each year respectively. The table I to V shows the first level EM clustering results.

By using the EM algorithm, no pre-defined number of clusters is needed for clustering model. For each year, the number of clusters is based on values of the diagnosed Diabetes Prevalence Rate (DPR) and the Uninsured Rate (UR) of the counties. The range of total number of clusters is from 6 to 9. The results show that the overall diabetes prevalence is increasing year by year, whereas the uninsured rate is in the deceasing trend other than from 2009 to 2010.

TABLE III
FIRST LEVEL EM CLUSTERING RESULT - 2011

	Clusters							
	0	1	2	3	4	5	6	Overall
No. of Counties	349	248	210	637	736	440	525	3142
DPR								
mean	7.69	9.29	11.025	11.028	8.87	7.33	10.81	9.52
std. dev.	0.83	0.67	1.29	2.44	1.05	1.11	1.16	2.05
UR								
mean	10.44	25.33	22.18	20.24	13.81	18.64	17.25	18.00
std. dev.	1.40	5.39	3.81	2.46	2.07	6.25	1.88	5.49

TABLE IV
FIRST LEVEL EM CLUSTERING RESULT - 2012

	Clusters					
	0	1	2	3	4	Overall
No. of Counties	752	70	335	1479	506	3142
Attribute - DPR						
mean	8.91	9.03	7.56	10.76	9.30	9.60
std. dev.	1.03	0.36	0.93	2.18	2.09	2.10
Attribute -UR						
mean	13.17	26.91	10.08	18.58	22.44	17.58
std. dev.	2.01	5.13	2.20	2.59	4.48	5.39

It is noticed that in each year there is a cluster (cluster 3 of 2009, cluster 8 of 2010, cluster 0 of 2011, cluster 2 of 2012 and cluster 3 of 2013) which has the lowest value of UR along with the lowest value or second lowest value of DPR. However, based on the values of the DPR and UR of these clusters, it is identified that although the UR values is consistent decreasing along the years, there is no consistent decreasing of the DPR values. The calculated correlation coefficient (using Eq. 1) between DPR and UR in these clusters of 2009 to 2013 are -0.14, -0.20, -0.11, -0.13 and 0.01 respectively. It indicates except 2013, there are negative correlations between the DPRs and URs for other years, but the relationships are gradually moving towards positive.

On the other end, the clusters (cluster 1 of 2009, cluster 7 of 2010, cluster 1 of 2011, cluster 1 of 2012 and cluster 1 of 2013) that have the highest value of UR do not have high DPR values, but median DPR values. The calculated correlation coefficient between DPR and UR are 0, -0.05, -0.01, 0.06 and -0.18 for year 2009 to 2013. These values indicate that for the counties in those clusters, there are no strong correlations between the DPRs and URs for years from 2009 to 2012. However, in 2013, it shows some negative relationship.

Lastly, it is also noticed that the clusters (cluster 4 of 2009, cluster 4 of 2010, cluster 3 of 2011, cluster 3 of 2012 and cluster 5 of 2013) that have highest value of DPR often have median value of UR. The calculated correlation coefficient between DPR and UR in these clusters are -0.28, -0.08, -0.07, 0.08 and 0.08 for year 2009 to 2013 respectively. This shows that correlations between the DPRs and URs are moving away from negative but towards positive.

TABLE V
FIRST LEVEL EM CLUSTERING RESULT - 2013

	Clusters						
	0	1	2	3	4	5	Overall
No. of Counties	102	98	831	372	735	1004	3142
Attribute - DPR							
mean	11.36	8.59	8.36	7.74	8.99	11.47	9.62
std. dev.	2.76	0.28	1.57	0.89	1.15	1.44	2.19
Attribute -UR							
mean	17.93	26.27	19.90	9.99	13.61	19.58	17.58
std. dev.	2.12	5.71	6.09	1.48	1.98	3.33	5.52

TABLE VI
TYPES OF CLUSTERS AND DESCRIPTION

Cluster Type	Color	Description	Year - Cluster
CT1	Dark Blue	Lowest UR (with lowest or second lowest DPR)	2009 - 3; 2010 - 8 2011 - 0; 2012 - 2 2013 - 3
CT2	Orange	Highest UR	2009 - 1; 2010 - 6 2011 - 1; 2012 - 1 2013 - 1
CT3	Purple	Highest DPR	2009 - 4; 2010 - 4 2011 - 3; 2012 - 3 2013 - 5

TABLE VII
SUMMARY OF STATES THAT HAVE MORE THAN 5 COUNTIES THAT WERE IN CT1 IN 2009 BUT MOVED IN 2010

	No. of counties	
State	2010 - 0	2010 - 1
Illinois	50	2
Iowa	21	10
Kansas	7	5
Michigan	11	0
Minnesota	8	3
Nebraska	11	8
Ohio	7	0
Virginia	8	2
Pennsylvania	18	0
Wisconsin	13	4

Based on the previous analysis, three types of clusters are summarized in table VI with cluster types, color on map, description and the specific clusters of each year that are included. Figures 2 to 6 show the distributions of counties in these three types of clusters on the U.S. map for the five years. Although the data of Hawaii and Alaska are included in the analysis, but they are not shown on the map.

Over the five years, the majority of the counties in CT1 are distributed in the central north region and some of them are in the New England area. The number of counties in CT1 decreased from 2009 to 2010 and stayed around 350. The counties that in CT1 of 2009 but not in CT1 of 2010 are further investigated. Table VII summarizes the states that have more than 5 counties that were in CT1 in 2009 but moved out of CT1 in 2010. It is found that many of them moved to cluster 0 of 2010 which has slightly higher in both URs and DPRs. A few of them moved to cluster 1 of 2010 which has much higher URs but similar DPRs.

The counties that are in CT2 mainly at the central south region, and many of them are in Texas. From year 2011 to 2012 the number of counties in CT2 reduced significantly. The notable change can be visualized in Texas. The findings show that these counties in Texas have moved from cluster 1

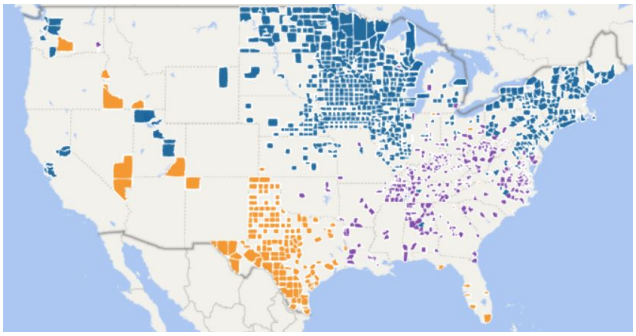

Fig. 2. Three Types of clusters distribution - 2009

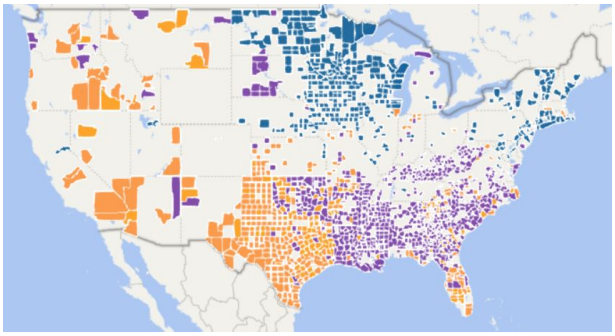

Fig. 3. Three Types of clusters distribution - 2010

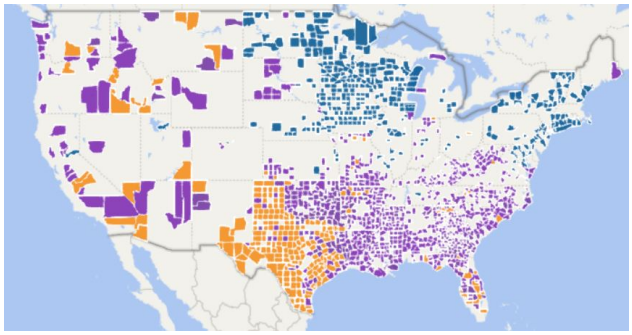

Fig. 4. Three Types of clusters distribution - 2011

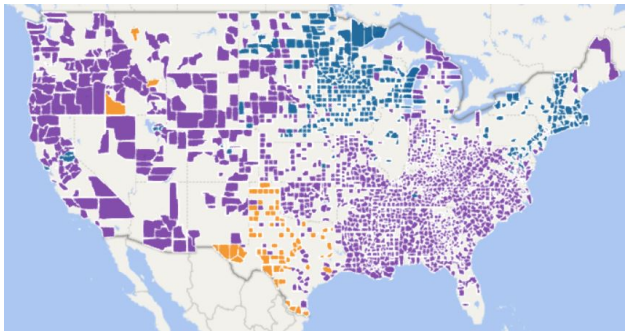

Fig. 5. Three Types of clusters distribution - 2012

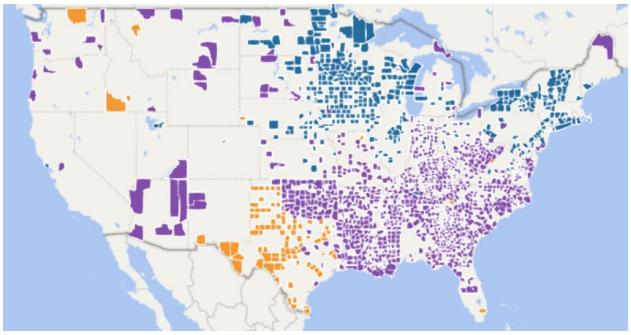

Fig. 6. Three Types of clusters distribution - 2013

of 2011 to cluster 4 of 2012. These movements indicates the although the URs of these counties have been reduced, the DPRs stays the same. In other words, increasing the health insurance rate has very minimum affect to the DPR for those areas over that time period. The possible reason could be that the diabetes prevention services were not covered under the insurance for that area or the awareness of those services was low.

The counties in CT3 have highest diabetes prevalence rate. The number of counties in CT3 have increased year by year from 2009 to 2012, but with a decrease from 2012 to 2013. Especially, the counties that are in the west and the central west regions have moved out of the CT3 in 2013. Table VIII summarizes the states that have more than 15 counties that were in CT3 in 2012 but not in CT3 in 2013. The results show that 386 counties have moved to cluster 2 of 2013. Among these counties, many are in California, Colorado, Missouri, and Kansas. These counties have the DRPs decreased without notable decreasing of the URs. There are 102 counties that have moved to cluster 0 of 2013. Among these counties, 31 are in Alabama. The majority of these counties have the URs slightly reduced with a slightly increasing of the DPRs. There are 118 counties have moved to cluster 4 of 2013. Many of them are in Indiana, Missouri and Virginia. These counties have significantly decreased URs along with some decreasing of DPRs.

B. Results of the second level EM and Discussion

The results of the first level EM clustering show that correlations of DPRs and URs of the counties can be similar or different for a given year. Counties' moving from one cluster to another demonstrates that the correlations between URs and DPRs over years are worth to investigate. In order to model the correlations over years, the second level EM clustering is built on top of the results of the first level EM clustering which is demonstrated in figure 1. Given that there are five consecutive years in this study, the size of the input vectors is 4. To avoid bias, after generating vectors for all the counties, the values are normalized to range [-1, 1].

The second level EM clustering results are given in table IX. The results show that most of clusters show mixed values of above 0 and below 0. This demonstrates that there is no strong positive or negative correlation between DPRs and URs over the years. Clusters 1, 2, 5 and 6 have three or all values

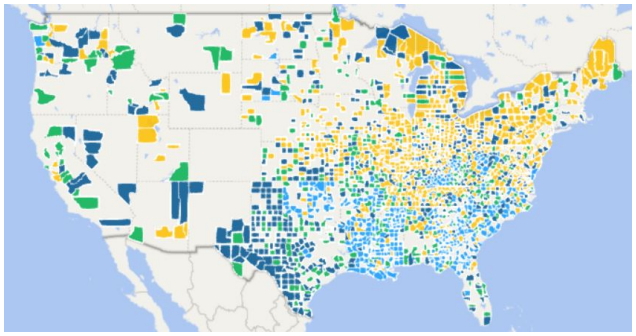

Fig. 7. Distributions of clusters 1 (Yellow), 2 (Dark blue) and 5 (Green), and 6 (Light blue) of second level EM clustering result

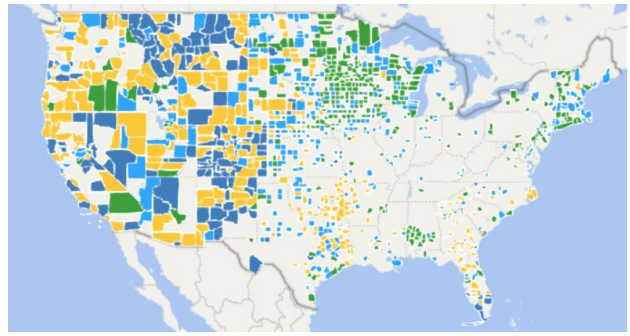

Fig. 8. Distributions of cluster 0 (Dark blue), 3 (Yellow), 4 (Green) and 7 (Light blue) of second level EM clustering result

TABLE VIII
SUMMARY OF STATES THAT HAVE MORE THAN 15 COUNTIES THAT WERE IN CT3 IN 2012 BUT MOVED IN 2013

State	# of counties		
	2013 - 0	2013 - 2	2013 - 4
Alabama	31	2	0
California	0	31	2
Colorado	0	20	1
Idaho	0	20	2
Indiana	0	8	11
Kansas	0	23	0
Kentucky	14	3	0
Michigan	0	6	11
Kansas	7	5	9
Missouri	1	38	14
Nebraska	0	22	0
North Carolina	0	23	3
Oregon	0	21	3
South Dakota	5	11	4
Texas	0	31	0
Virginia	1	21	21
Washington	0	14	2

TABLE IX
SECOND LEVEL EM CLUSTERING RESULT

	Clusters							
	0	1	2	3	4	5	6	7
No. of Counties	163	733	573	307	322	327	391	326
$PLA.slope$ - 2012 to 2013								
mean	0.46	0.21	0.52	-0.04	-0.92	-0.74	0.40	0.36
std. dev.	0.05	0.08	0.31	0.05	0.01	0.20	0.01	0.29
$PLA.slope$ - 2011 to 2012								
mean	0.65	-0.03	0.07	-0.99	0.15	0.05	0.07	0.01
std. dev.	0.08	0.02	0.26	0.01	0.03	0.07	0.04	0.39
$PLA.slope$ - 2010 to 2011								
mean	-0.39	0.93	0.52	-0.03	-0.31	0.31	0.18	-0.63
std. dev.	0.13	0.03	0.22	0.03	0.04	0.32	0.18	0.23
$PLA.slope$ - 2009 to 2010								
mean	-0.42	0.28	0.31	0.02	0.03	0.25	0.87	-0.21
std. dev.	0.08	0.10	0.39	0.09	0.17	0.37	0.01	0.37

above 0. This implies that over the years, the DPRs and URs of these counties are more in positive correlations. These counties distribute in the central south and eastern region as demonstrated in figure 7. On the other side, clusters 0, 3, 4 and 7 have two or more $PLA.slope$ values below 0. It indicates that the DPRs and URs of those areas are more in negative correlations. Figure 8 shows these counties distribute in the central north and western region, which are different from the counties are more in positive correlations over the years.

V. CONCLUSION, DISCUSSION AND FUTURE WORK

Through this research, geographic patterns of the correlations between uninsured rates and diabetes prevalence rates at a large scale in U.S. have been examined. The first level EM clustering results have shown that over the five years (2009 - 2013), the correlations in these two rates of the some clusters are gradually moving from negative towards positive. Based on the distributions of the clusters, there is no significantly change on distributions of the clusters that have highest UR, or highest DPR, or lowest UR in U.S.. The counties that had lowest DPR along with the lowest UR are located in the Central North and New England regions. The counties that had highest UR with median diabetes prevalence are in the Texas and Central South region. The counties that had highest DPR with median uninsured rate are in the Eastern and South East region. The number of the counties in these different clusters have changed between years. Between two consecutive years, although some counties have the uninsured rate stayed the same or even slightly increased, the diabetes prevalence rate decreased. The possible reasons could be that some of the counties have the diabetes prevalence rate under control through other strategies instead of relying on the health insurance. On the other end, some of the counties have the diabetes prevalence rate decreased along with the decreasing of uninsured rate. It implies that the diabetes preventive services under the insurance do have some impact on the diabetes prevalence. However, the usage of the preventive services need be further investigated and evaluated. The second level EM clustering results demonstrate the correlations between DPRs and URs over the consecutive years. For some regions, the correlations over time are negative, whereas for some other regions, they are positive. For the counties that the correlations over time are negative, the reason behind might be that some insurances might cover the diabetes preventive services for certain population but not the majorities of the population who might develop diabetes in that region. The other reason could be the differences in the awareness of the diabetes preventive

services that are covered under the health insurance policies [12].

Overall, the results suggest that future study need to be done on the uninsured rate combining with other factors for diabetes prevalence rate for different geographic regions. Combining with other factors might shed a light that whether the insurance need to be customized based on the population needs of a specific geographic region. In this study, the analysis of the correlations between the uninsured and diabetes prevalence rates over years was based on aggregation of two consecutive years. Future work can extend it to evaluate the continuous correlations over consecutive years.

REFERENCES

[1] D. R. Whiting, L. Guariguata, C. Weil, J. Shaw, IDF Diabetes Atlas: Global estimates diabetes for 2011 andd 2030, Diabetes Research and Clinical Practice, 2011, pp. 311–321

[2] J. E. Shaw, R. A. Sicree, P. Z. Zimmet, Global estimates of the prevalence of diabetes for 2010 and 2030, Diabetes Research and Clinical Practice, 2010, pp. 4–14

[3] A. M. McBean, S. Li, D. T. Gilbertson, A. J. Collins, Differences in diabetes prevalence, incidence, and mortality among the elderly of four racial/ethnic groups: whites, blacks, hispanics, and asians, Diabetes Care, 2004, 27(10), pp. 2317–2324

[4] H. King, R. E. Aubert, W. H. Herman, Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections, Diabetes Care, 1998, 21(9), pp. 1414–1431

[5] X. Zhang, D. E. Williams, K. M. Bullard, L. E. Barker, E. W. Gregg, A. L. Albright, G. L. Beckles, G. Imperatore, Access to Health Care and Control of ABCs of Diabetes, Epidemiology/Health Services Research, 2012, 35, pp 1566–1571

[6] J. E. DeVoe, C. J. Tillotson, L. S. Wallace, Usual Source of Care as a Health Insurance Substitute for U.S. Adults With Diabetes?, Diabetes Care, 2009, 32(6), pp. 983–989

[7] J. Hill, M. Nielsen, and M. H. Fox, Understanding the Social Factors That Contribute to Diabetes: A Means to Informing Health Care and Social Policies for Chronically III, The Permanente Journal, 2013, 17(2), pp. 67–72

[8] L. W. Pickle and Y. Su, Within-State Geographic Patterns of Health Insurance Coverage and Health Risk Factors in the United States, American Journal of Preventive Medicine, 2002, 22(2), pp. 75–83

[9] A. H. Mokdad, M. M. Engelgau, E. S. Ford, F. Vinicor, B. A. Bowman, J. S. Marks, and D. E. Nelson, Diabetes Trends in the U.S.: 19901998, Epidemiology/Health Services/Psychosocial Research, 2000, 23(9), pp. 1278–1282

[10] About the Law: Preventive Care, U.S. Department of Health & Human Services, https://www.hhs.gov/healthcare/about-the-law/preventive-care/index.html

[11] J. Aaron Hipp and N. Chalise, Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009 - 2010, Preventing chronic disease, 2015, 12

[12] Preventive Services Covered Under the Affordable Care Act, National Conference of State Legislatures, http://www.ncsl.org/research/health/american-health-benefit-exchanges-b.aspx#15

[13] CDC County Data Indicators - Diagnosed Diabetes Prevalence, https://www.cdc.gov/diabetes/data/countydata/countydataindicators.html

[14] Small Area Health Insurance Estimates (SAHIE), https://www.census.gov/did/www/sahie/data/index.html

[15] L. E. Barker, K. K. Kirtland, E. W. Gregg, L. S. Geiss, T. J. Thompson, Geographic Distribution of Diagnosed Diabetes in the U.S., American Journal of Preventive Medicine, 2011, 40(4), pp. 434–439

[16] A. Michimi, M. C. Wimberly, Spatial Patterns of Obesity and Associated Risk Factors in the Conterminous U.S., American Journal of Preventive Medicine, 2010, 39(2), pp. 1–10

[17] H. King, M. Rewers, and Who Ad Hoc Diabetes Reporting Group, Global Estimates for Prevalence of Diabetes Mellitus and Impaired Glucose Tolerance in Adults, Diabetes Care, 1993, 16(1), pp. 157–177

[18] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, 1977, 39(1), pp. 1-38

[19] A. Khasawneh, S. A.Alvarez, C. Ruiz, S. Misra and M. Moonis, Discovery of sleep composition types using Expectation-Maximization, Proceedings of IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), 2010, pp. 26–31

[20] F. Sufi and I. Khalil, Diagnosis of Cardiovascular Abnormalities From Compressed ECG: A Data Mining-Based Approach, IEEE Transactions on Information Technology in Biomedicine, 2011, 15(1), pp. 33-39

[21] R Sridharan, Gaussian mixture models and the EM algorithm,

[22] Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[23] M. Kontaki, A. N. Papadopoulos, and Y. Manolopoulos, Continuous Trend-Based Clustering in Data Streams, Proceedings of 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK2008), 2008, pp. 251–262

[24] Diabetes 2014 Report Card, https://www.cdc.gov/diabetes/pdfs/library/diabetesreportcard2014.pdf