USING CLUSTERING SOFTWARE FOR EXPLORING SPATIAL AND TEMPORAL PATTERNS IN NON-COMMUNICABLE DISEASES

Mariana Nagy
"Aurel Vlaicu" University of Arad Romania
Department of Mathematics and Computer Science / Head of Department
Dana Negru
Arad, Public Health Department

Abstract

The paper deals with clustering methods that can be used for detecting spatial and temporal patterns in large databases. The basics of some common methods are presented. The authors use the cancer records in Arad County for the last 55 years to apply methods for cluster detection. The statistical approach reveals the existence of several spatial and one temporal cluster. Research will be continued by further analysis in order to refine the results and help authorities in defining correct health policies.

Keywords: Cancer statistics, clustering methods, statistical software, disease clusters

Introduction

Worldwide databases are set-up in order to register data on different diseases. In the last two centuries, mainly since the time of the London cholera outbreak, scientists are more and more motivated to undertake relevant studies on the available data, to identify risk factors and to provide authorities scientific support for health policies.

The most usually, diseases are recorded through their location, time occurrence, personal data of patients and the medical evolution of the cases. Other considered data are the disease onset, date of diagnosis, age and gender of patients.

One of the most relevant analysis is related to the aggregation of cases in space, in time or both in space and time. Disease clusters appear when more cases then normally expected are identified in a region or/and in a certain time and in a certain population. Once the areas with high

prevalence of a certain disease are identified and studied, in the next step, the associated hazards will be determined and appropriate action will be taken in order to decrease the risk factors.

However, causes of disease clusters may not be clearly identified, spatial locations of cases often offer at least an indirect estimation for exposure to a risk factor (Thomas & Carlin, 2003). Additionally, the precise date of disease onset is often unavailable and may be only estimated with the date of diagnosis. Being based on incomplete knowledge, clustering methods can be used mainly for identifying patterns and generating hypothesis that have to be confirmed latter.

Clustering methods

From a statistical point of view, a cluster is defined as a grouping or nesting of phenomena or events that occur in a defined geographic proximity, in a limited period of time or both, in a certain population. Depending on these, clusters can be spatial, temporal or spatio-temporal. For each cluster type, specific statistical methods can be used in order to reveal the existence of clusters and for studying its characteristics.

Spatial clusters

Spatial cluster analysis plays an important role in quantifying geographic variation patterns. It is commonly used in disease surveillance, spatial epidemiology, population genetics, landscape ecology, crime analysis and many other fields. (Jaquez, 2008)

Spatial cluster detection methods explore the area of the events and evaluate whether there exist particular locations where events tend to aggregate. The detection methods are general or focused (Besag & Newell, 1991), while the general methods can be further classified in global or local ones (ClusterSeer, 2012). Subsequently, clusters can be global, local or focused:

- Global methods detect clustering throughout the study area regardless of their specific locations or spatial extent. Examples are: Kulldorff scan statistics, Ripley's K-function, Besag and Newell's R Method.
- Local methods detect clustering limited to geographically restricted areas within the study. Examples are: Besag and Newell's L method, Turnbull's method, Local Moran test.
- Focused methods detect clustering around a specific location. Candidate locations can be represented by the presence of a risk factor. Examples are: Diggle's Method, Bithell's method, Score test.

One of the most relevant detection methods, frequently used for analyzing disease databases is Besag-Newell’s method. The method can be applied both at global (R) and local (L) assessment, at individual level or as a

region centered group-level technique. Being a basic statistical test a null hypothesis and a work hypothesis must be defined. The null hypothesis H0: the cases within the studied area comply with Poisson distribution, assumed as normal. The work hypothesis that contradicts H0: there are areas where the number of cases exceeds that predicted by the Poisson distribution (considered normal). The spatial areas that present higher densities than the normal occurrence expected is considered a region cluster.

In order to test H0, around each region, a circular window is considered, sized to include a specified number of cases k. Then, the population inside the window is compared to the normal (Poisson) distribution. If the frequency of cases is higher than the expected (Poisson) one, H0 is rejected, the work hypothesis is accepted and the region is considered a cluster. If the method is applied for an area where the density of population is not particularly low, the results are similar to those obtained by other methods, like Kulldorff spatial scan. (Costa, Assunção, 2005)

Temporal clusters

Temporal cluster analysis is used for identifying and investigating whether there exist a period of time where cases present a higher density and which is the temporal pattern for the evolution of the number of cases in a certain geographical area.

The temporal clustering methods can be applied both to data at group level and individual level, being used to evaluate disease frequency or case counts in a single or in multiple time series. The most known algorithms for determining and analyzing temporal clusters are: Dat's Method, Ederer-Myers-Mantel Method, Empty Cells, Grimson's Method, Larsen's Method, Levin and Kline's modified Cusum, Wallenstein's Scan.

One of the most intuitive methods for monitoring the pattern of cases over time is Levin and Kline's modified Cusum. The database will contain counts of cases and the total population at risk. The term CuSum stands for Cumulative Sum and was used by Page to analyze the variation of a variable relative to a baseline value. (Page, 1961). The method was then modified by Levin and Kline and adapted to the temporal analysis of epidemiologic data.

As for any statistical test a null hypothesis and a work hypothesis are defined. The null hypothesis H0: the cases are spread at a homogeneous rate over time. The work hypothesis that contradicts H0: there are moments in time where the rate of cases is temporarily elevated. (ClusterSeer, 2012)

Testing the statistics consists of calculating the modified CuSum value for every time sequence of fixed length, according to (1), then comparing the largest CuSum values to Monte Carlo distribution. (Levin & Kline, 1985)

$$W_t(r) = \max(0, W_{t-1} + Y_t - r) \qquad (1)$$

where:

- t = 1, 2, … and $W_0 = 0$;
- Y_t is the case count in the interval t ;
- W_{t-1} is the CuSum for the interval (t-1)
- r is the reference value, the baseline level that gives the sensitivity of the model and is calculated using the Relative risk ω, the average risk λ_0 and the population at risk n, according to (2):

$$r = n\frac{\lambda_0(\omega-1)}{\log\omega} \qquad (2)$$

Spatio-temporal clusters

Spatio-temporal methods are focused to space – time interaction and detect case clusters in space that depend on the time period.

The spatio-temporal clustering methods can be applied both to data at group level and individual level. The most known algorithms for determining and analyzing temporal clusters are: Direction Method, Kulldorff's Spatio-Temporal Scan, Jacquez's k-Nearest Neighbor Method, Grimson's Method, Knox's Method, Mantel's Method.

Kulldorff's scan method can be used for detecting spatio-temporal clusters without beforehand requiring the specification of spatial or temporal extent. There are two Kulldorff models, based on Poisson and Bernoulli distributions: the first one is preferable for continuous variables or case counts where exposure is important while the second one gives better results in questions of yes/no binary counts (Kulldorff, 1997).

The null hypothesis H0 defines the null spatial model as an inhomogeneous Poisson point process with an intensity that is proportional to the population-at-risk. The work hypothesis H1 states that in some locations in the multidimensional space, the number of cases exceeds the predicted number under the null model. (ClusterSeer, 2012)

In order to test H0, a cylindrical space – time window is considered for each region at a particular time. The window is expanded to include neighboring regions and time intervals until it reaches a maximum size of 50% of the average population at risk and 50% of the span period. The number of cases within the window is evaluated at each window size and compared to the null spatial model at that particular time. The P value is obtained through Monte Carlo randomization. If the null hypothesis is rejected, at least one spatio-temporal cluster can be identified in the data. (ClusterSeer, 2012)

This type of analysis is suitable mainly for historical data for which a Poisson distribution is considered normal.

Cluster detection in the cancer database of Arad County

The following examples perform cluster detection in the overall cancer database of Arad County. Cancer is recorded systematically in Romania since 1974, while County Cancer Registry was conducted electronically after 2005 in Arad and contains a number of 19,730 cases (*Authors*, 2014). The data are gathered from different sources and are aggregated at the public Heath Department of Arad County. The data cover a time period from 1959 to 2013 for all the 75 regions in the county. The cancer types are clearly identified for each case in the database, but for the present clustering exercise only the total number of cases is considered for each region and period.

The computer instrument used for cluster detection is ClusterSeer, version 2012, trademark of BioMedware. This is a geospatial research software that offers data visualization tools and statistical tools to explore spatial and temporal patterns of disease. (ClusterSeer, 2012)

Detecting Spatial clusters

For spatial analysis, the study considers all the new cancer cases recorded in 2008 in Arad County. The county is divided in 75 regions – potential clusters.

The data was analyzed applying Besag-Newell's method as a region centered group analysis, looking for both local and regional clusters. The data file includes the 75 regions, the geographical coordinates of the region centers, the number of new cases recorded in 2008 and the population at risk for each region. The file is set-up in text format, according to the requests of ClusterSeer software. The size of the circular window (k) was established by successive attempts.

The following hypothesis are stated:

H0: In Arad County, cancer cases in 2008 are normally distributed, according to Poisson statistics.

H1: There are areas with higher case densities that those predicted by Poisson statistics.

By using the ClusterSeer implementation for the method, three type of results are available: a Session Log, a Map and a Histogram/Monte Carlo distribution. The chosen confidence level for the analysis was α=5%.

The results are presented in Figure 1 (Session Log) and Figure 2 (Map). The program identified 2 local clusters for cut-off size k=21, located as shown on the map. Thus, Felnac and Fintinele are local clusters for all tumor disease in 2008, while a regional cluster might be located in each of the remaining 73 centroid region.

```
Besag & Newell's Method 10/08/2014  09:43 PM
************************

Cluster size to detect = 21
Alpha level = 0.05

Centering  Local Disease   (1) Test  Upper-tail
Region     Frequency      Statistic  P-value
================================================
FELNAC     0.00713128             1  0.010161
FINTINELE  0.00606411             1  0.030643

Total number of significant local clusters (r) = 2
Expected size of r under null hypothesis = 0.325600
Upper-tail P-value for r from Monte Carlo distribution (999 simulation runs)= 0.044000
```

Figure 1 – Results for seeking clusters with k = 21 and α=0.05

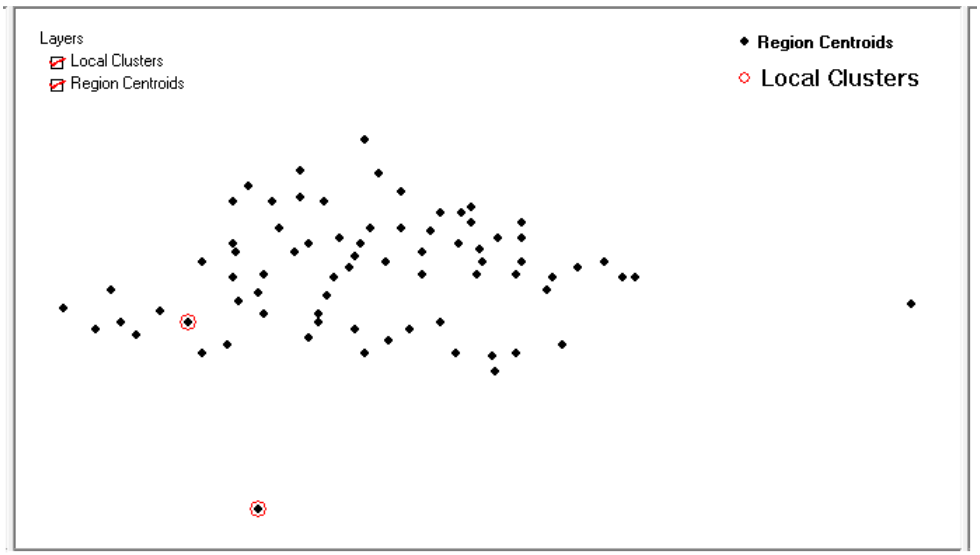

Figure 2 – Map for local and regional clusters

Further studies will consider a segregation of the population by age *(Authors*, 2014), by sex and by cancer type and will be extended for other periods corroborating those with spatial or spatio-temporal studies.

Detecting temporal clusters

For identifying temporal clusters, one of the most suitable methods is Levin and Kline's modified Cusum. The data files include the number of new cases yearly recorded and the population at risk. Because exact data on the population are not available, a linear interpolation is accepted. The following hypothesis are stated:

H0: In Arad County, cancer new cases are recorded according to a uniform time distribution.

H1: There are time periods with a higher case frequency which can be considered time clusters.

A preliminary graphical analysis, shown in Figure 3, suggests a peak of cases between 2006 and 2012.

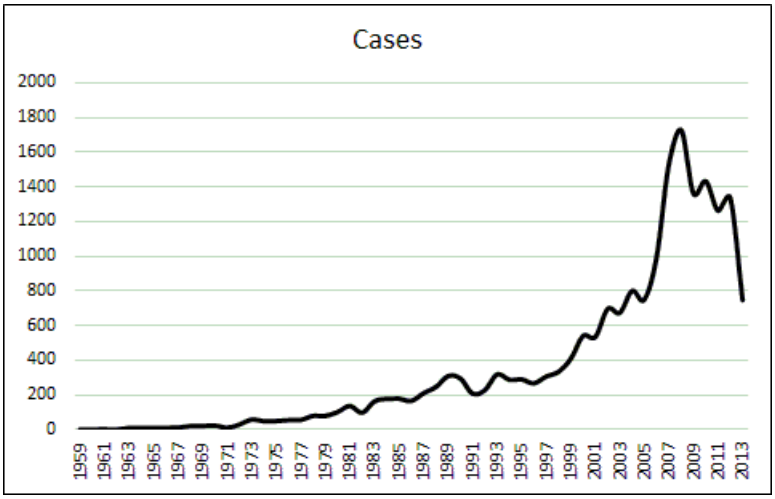

Figure 3 – Evolution of cases for 1959 – 2013

Considering new cases recorded in 2007 - 2011 as the exposed group and the whole sample as the control group, the Relative risk is computed. Thus, for the peak period, the Relative risk is ω = 2.7. Using this value, Levin and Kline modified CuSum analysis is run with ClusterSeer. The results are presented in Figure 4 (Session Log) and Figure 5 (Plot).

```
Dataset source:
D:\AnalizaClusterseer\LevinKline\CaseT.txt  09/11/2014  10:38 PM
D:\AnalizaClusterseer\LevinKline\CensusT.txt  09/11/2014  10:51 PM
Census data was used to estimate population-at-risk size using linear extrapolation
Total number of time intervals = 55
Study period span = 1959 - 2013
Average annual disease frequency = 0.000940989


Levin and Kline's Modified Cumulative Sum Procedure
**************************************************

Relative risk used for analysis = 2.73000
Alpha level = 0.05
Number of Monte Carlo simulations performed = 999

      Time  Interval  Disease       CuSum         Upper-tail
    Period  Sequence  Frequency     Statistic     P-value
============================================================
      2008        50  0.00394530    93.3236       0.001000
```

Figure 4 – Results for seeking time clusters in Arad County

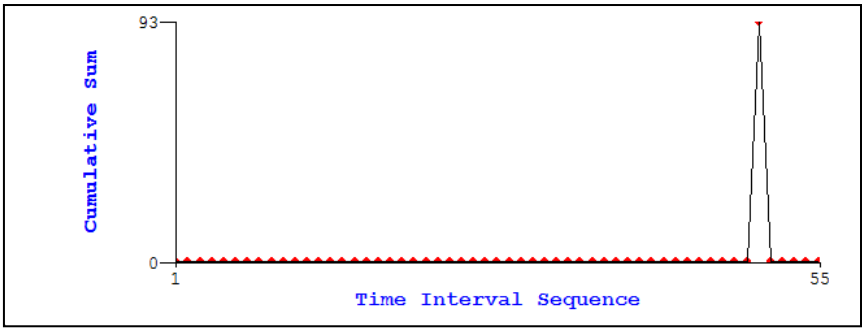

Figure 5 – Plot for the modified CuSum

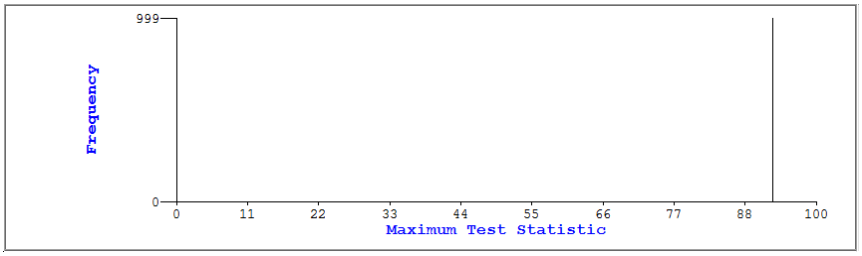

Figure 6 – Result of Monte Carlo simulation

The program identified one time cluster for the 50th time period, corresponding to 2008. The Plot highlights the cluster much more precisely than the Chart in Figure 3. The same result is given by the maximum in Monte Carlo simulation (Figure 6).

The sensitivity of the analysis is given by the value of r. By computing it according to (2), the baseline is established to r = 26.54. Thus, the cumulative sum considers only periods where the number of cases increased at least with 26.54.

Detecting spatio-temporal clusters

Using the same database, a more advanced analysis can be performed. In order to find a relationship between the space and time clusters, a Kulldorff spatio-temporal scan is applied. The data files include the same 75 centroid regions and their coordinates, the new cases recorded each year in each region and the population at risk. The following hypothesis are stated:

H0: The new cases recorded in Arad County are distributed according a Poisson point process, uniform in time and proportional to the population at risk.

H1: There can be identified space-time areas where the number of new cases exceeds those predicted.

Using ClusterSeer, Kulldorff's spatio-temporal scan is run. Firstly, clusters are searched in the whole county (75 regions), including all 19,730

cases and for the whole time period (1959 – 2013). The implicit settings of the program are used: maximum spatial population radius analyzed (50% of total population) = 225601, maximum temporal span analyzed (50% of study period span) = 27 years, the number of Monte Carlo simulations performed = 999.

The analysis revealed one cluster for 2000 - 2013 time span, centered in Iratosu and including Sofronea, Curtici, Macea, Arad, Zimandu-Nou, Livada, Felnac, Pecica, Graniceri and Vladimirescu. The annual disease frequency is 0.00263227 while the likelihood ration of cases in the cluster is 5251.15.

A second Kulldorf scan was run, restricting the spatial search area only to the inside of the firstly identified cluster. Two space-time clusters were found:

- The first most likely cluster was detected for 2006 – 2012 time span, centered in Graniceri, including Macea, Curtici, Iratosu, Sofronea, Zimandu-Nou and Livada, with an annual disease frequency of 0.00283841 and a likelihood ration of cases of 275.89.
- The second most likely cluster was determined for 2005 – 2013 time span, being centered in Felnac and including Pecica too. The annual disease frequency is 0.00289380, while the likelihood ration of cases in the cluster is 204.498.

A third analysis searched clusters considering only 2004 – 2013 time span and the previously restricted geographical area. Applying the space-time limitation is justified mainly none of the previous space, time or spatio-temporal methods identified clusters outside this multidimensional area. Furthermore, for 1959 – 1989 time span, the available data might be less reliable due to an inaccurate recording.

Thus, the Kulldorff scan considering 11 regions and the last decade detected two spatio-temporal clusters:

- The first most likely cluster is identified in Felnac, for 2007 – 2009 time span. The annual disease frequency is 0.00545171 and the likelihood ration of new cases is 7.03413.
- The second most likely spatio-temporal cluster is centered in Zimandu-Nou, including Livada, Sofronea and Curtici. The time span is 2008, the annual disease frequency: 0.00461906, while the likelihood ratio of cases is 6.63902.

The results given by Clusterseer software are presented in Figure 7 (map), Figure 8 (plot) and Figure 9 (histogram / Monte Carlo Simulation). The two maximum statistics obtained in Monte Carlo simulation correspond to the likelihood ratio of cases in the two detected clusters. For both clusters, the temporal maximum is around the year 2008.

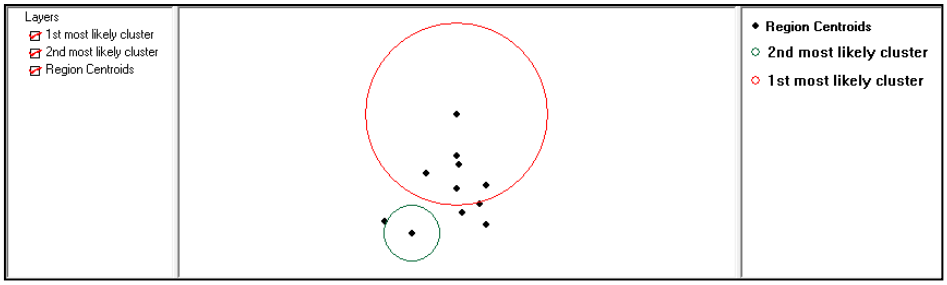

Figure 7 – Map for locating the spatio-temporal clusters

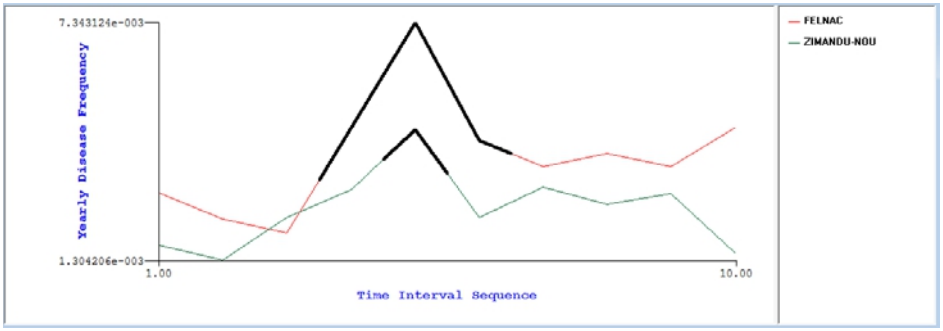

Figure 8 – The time plot for the two detected spatio-temporal clusters

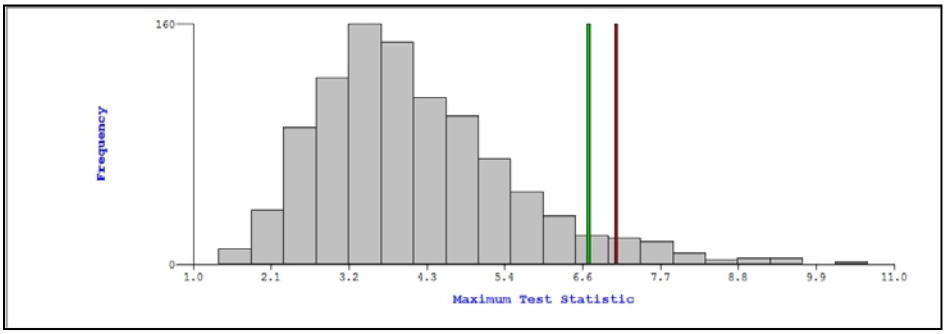

Figure 9 - The histogram and Monte Carlo simulation

The spatio-temporal clusters found show that the previously detected spatial and temporal clusters are not independent: in the cancer database of Arad County there is a strong space-time connection proved by the Kulldorff scan mehod.

Conclusion

Among the available statistical methods, clustering is particularly useful for identifying space and time patterns in large data sets. ClusterSeer - geospatial research software, offers an easy to use implementation for the clustering algorithms. It includes spatial, temporal or spatio-temporal methods, applied at global or local level. Applied on the cancer database of

Arad County, the results revealed the existence of space-time areas with higher disease prevalence than expected. All of the presented statistics detected at least one case aggregation in space and time, centered in Felnac around year 2008.

Further research are developed for in-deep analysis of historical data and correlations with environmental, behavioral and work health hazards, in order to help authorities to establish and apply better public health policies.

Acknowledgements

The authors express their gratitude to the Arad Public Health Department and the National Statistics Office / Arad County for making available the data used in the present research.

References:

Besag, J., Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154, 143-155;

Costa, M.A., Assunção, R.M. (2005). A fair comparison between the spatial scan and the Besag–Newell Disease clustering tests. *Environmental and Ecological Statistics*, 12(3), 301-319;

Jacquez, GM. (2008). Spatial Cluster Analysis. *The Handbook of Geographic Information Science*, Blackwell Publishing, 395-416, available at http://www.biomedware.com /files/jacquez_ch22_preprint.pdf (download: august 2014);

Kulldorff, M. (1997). A spatial scan statistic. *Statistics–Theory and Methods*, 26, 1481-96;

Levin, B., Kline, L. (1985). The cusum test of homogeneity with an application in spontaneous abortion epidemiology, *Statistics in medicine*, 4(4), 469-488;

Authors. (2014).

Page, E.S. (1961). Cumulative sum charts, *Techonometrics*, 3, 1-9;

Thomas, A., Carlin, B. (2003). Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. *Statistics in Medicine*, 22, 113-27;

***ClusterSeer (2012), User Manual, v.2.5, BioMedware, Inc.