Cluster Analyses of Association of Weather, Daily Factors and Emergent Medical Conditions

Jasmin Malkić[1], Nermin Sarajlić[2], Barbara U. R. Smrke[3] and Dragica Smrke[4]

[1] University Hospital, Diagnostics, Anesthesia and Technology Division, Uppsala, Sweden
[2] University of Tuzla, Faculty of Electrical Engineering, Tuzla, Bosnia and Herzegovina
[3] University Medical Center, Department of Neurosurgery, Ljubljana, Slovenia
[4] University Medical Center, Department of Traumatology, Ljubljana, Slovenia

A B S T R A C T

The goal of this study was to evaluate associations between the meteorological conditions and the number of emergency cases for five distinctive causes of dispatch groups reported to SOS dispatch centre in Uppsala, Sweden. Center's responsibility include alerting to 17 ambulances in whole Uppsala County, area of 8,209 km² with around 320,000 inhabitants representing the target patient group. Source of the medical data for this study is the database of dispatch data for the year of 2009, while the metrological data have been provided from Uppsala University Department of Earth Sciences yearly weather report. Medical and meteorological data were summoned into the unified data space where each point represents a day with its weather parameters and dispatch cause group cardinality. DBSCAN data mining algorithm was implemented to five distinctive groups of dispatch causes after the data spaces have gone through the variance adjustment and the principal component analyses. As the result, several point clusters were discovered in each of the examined data spaces indicating the distinctive conditions regarding the weather and daily cardinality of the dispatch cause, as well as the associations between these two. Most interesting finding is that specific type of winter weather formed a cluster only around the days with the high count of breathing difficulties, while one of the summer weather clusters made similar association with the days with low number of cases. Findings were confirmed by confidence level estimation based on signal to noise ratio for the observed data points.

***Key words**: emergency department visit, primary health care, meteorology, data mining, Uppsala*

Introduction

It is a well documented fact that a number of the admissions in medical care in general follow its seasonal variations, especially when it comes to the asthma, allergies, and respiratory illness[1]. In order to show how these emergency cases quantities fluctuate under the different weather conditions, the analysis is conducted for the region of Uppsala, Sweden. This study was undertaken to further our understanding about whether certain meteorological conditions influence the number of cases reported to Uppsala SOS dispatch central. The analyses were not undertaken strictly to evaluate whether weather is statistically related to any of observed illness symptoms, as such procedure would require consideration of additional risk factors for each cause of dispatch group[2–5] (e.g. air pollution or outdoor aeroallergens levels for the breathing difficulties). Rather, the analyses apply several statistical methods to prepare the collected data in order to model certain data points used as material in the process of data mining. Therefore, the main objective was to examine the association between several meteorological conditions and emergency case quantities, and to determine could the weather be a useful indicator of the patient volumes as well. The escalation of medical costs, the limitations of related resources, and the necessity to provide cost-effective medical care underscore the value of forecasting patient volume to delivery of health care in general[6–8]. Furthermore, an improved understanding about how certain meteorological conditions affect the amount of emergency dispatch causes can facilitate medical intervention and behavior changes that could help handle these conditions.

Received for publication July 9, 2010

Materials and Methods

Source of the dispatch data for this study is the Uppsala SOS dispatch central database for the year of 2009. Since its introduction in the emergency service, the Zenit technical platform provides the comprehensive database of the emergency events covering not only patient's medical but case's geographical (GPS coordinates), temporal (time and the duration of the different event phases) and resource (e.g. ambulance transport) data as well. This study makes use of daily number of the cases sorted by five most frequent groups of dispatch causes, which would be:

- breathing difficulties,
- symptoms with suspicion of stroke
- chest pain and suspicion to heart disease
- headache or dizziness and
- abdominal or urinary tract symptoms.

Target patient group for the study represents the whole population of Uppsala County with around 320,000 inhabitants, no matter the age or gender, so data from the referential database satisfy the simple random sampling criteria. Over the whole year of 2009 (365 days) there was a total of 36,159 healthcare-related emergency cases reported to the service that related to the certain symptoms. Total of 29,922 cases were registered as an »ambulance event«, while those connected with illness symptoms make 87.99% of all the cases sorted into 26 groups by dispatch centre personal qualified to receive an emergency telephone call. It should be noted that this personal work without direct support from a medical staff. As they do not necessarily possess the medical education, they are not able to set the diagnoses, but merely to suspect it or notice the symptoms. Due to the very nature of emergency service, some dispatch cause groups are registered notably more often than the others, so roughly every third case belong to one of the five most frequent groups which, together with different kinds of accidents, makes about half of all the cases (Figure 1). In addition, all the ambulance events are classified into one of the four priority groups (Figure 2). Herein, the study is based on total of 10,084 cases connected with the five most frequent dispatch cause group in 2009. Besides being most frequent, a relation with the weather conditions has already been reported to most of these dispatch cause groups in other parts of the world and different climate conditions.

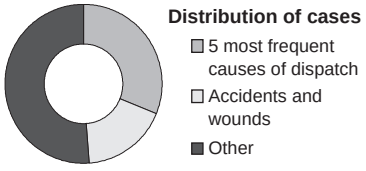

Fig. 1. Distribution of cases in Uppsala SOS dispatch centre.

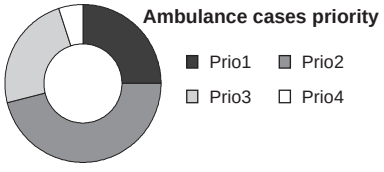

Fig. 2. Distribution of the priority groups within the Uppsala SOS Dispatch centre's healthcare-related cases.

Source of the meteorological data are Uppsala University Department of Earth Sciences daily weather reports from Celsius weather station, and the parameters taken into the consideration are:

- temperature difference (minimum to maximum daily value),
- mean air pressure (daily values),
- precipitation (daily values),
- mean relative humidity (daily values) and
- mean wind speed (daily values).

Uppsala County covers the area of 8,209 km^2 with 39 inhabitants per square kilometer, while the overwhelming majority of target population resides within a 50 km of the weather observing site.

The referential database

Following database fields are used:

- CaseNumber (unique case identification),
- Created (exact time when the dispatch centre became aware of the case) and
- CaseIndexNameX (cause of dispatch classification in total of 26 primary groups)

By the series of SQL queries, case data are transformed into the daily quantities of cases sorted by five most frequent cause of dispatch groups over the whole year, which gives the five vectors of 365 non-negative number values representing the daily case quantities.

The weather models

Region of Uppsala is known by its Scandinavian maritime climate dictated by extremely low altitude differences and the proximity of the Baltic Sea. In order to relate weather with emergency cases, the weather itself must be quantized. The weather data for each day are integrated with the daily case quantities to form the six dimensional points of the »weather space« with five dimensions for the weather parameters and one additional for the observed cause of dispatch group daily quantity. Every point of this space represents a day of the year with its weather conditions and case frequency. Purpose of such organization is at first hand to distinct the days within the observed period by their respective weather parameters and then to make a finer distinction between the days with fewer and higher number of emergency cases by the cause of dispatch group. Such »type of weather« models are well recognized as suitable to be combined with case frequency for the purpose of general health care analyses[9].

Statistics

Analyses of the six dimensional data space are made for each most frequent dispatch cause group separately. Essentially, it is the problem of finding the data clusters

in a spatial database and the point in space is represented by the particular dispatch cause group cardinality and five meteorological parameters. The clusters are formed out of 365 six-dimensional data points representing the dates of the year. The data mining method choice is density-based algorithm DBSCAN since it is, compared to the other similar clustering algorithms, much more robust considering the cluster shape and possible noise points[10,11]. Before DBSCAN can actually be applied the data space need to go through the process of certain statistical conditioning, expressed through the distance and variance normalization. Though different methods are used, there is still a well recognized common need for certain data preparation, before the actual statistical processing can be applied within a medical research[12]. In this case, the standard deviances for each data dimension (meteorological attributes or dispatch cause cardinality) were computed and compared to each other, so that no attribute could overshadow the significance of the others, which increases the reliability of point distance measuring. Although it is not the case with theirs variance and order of magnitude, physical units of the weather parameters are irrelevant for the cluster analyses, so the weather parameters were treated strictly as scalars. To find an orthogonal axles of the data space, the method of principal component analyses (PCA) is applied. Using the principal components as the axles of the data space it is possible to visualize the data points, at least in three visible, statistically most significant dimensions, and get the rough picture of the number of clusters and the presence of the noise, which helps to estimate the DBSCAN algorithm parameters.

As the next step in data conditioning cleaning out the noise points was done through determining the optimal points distance from their neighbor points. Point distance was calculated as the Euclidian distance in six-dimensional space. Points that have five of their closest neighbors within the optimal distance remained in consideration to form the clusters (signal), while others were considered as noise. To find the optimal distances (ε parameters), the 5th closest neighbor point distance graph was made for each dispatch cause group separately (meaning that MinPts parameter is set to the value of 5). The impact of optimal distance value choice is not limited to the clusters number and shape only, but to the statistical confidence in the analyses results as well. Expressed in words, the confidence is the ratio of the magnitude of the signal to the magnitude of the noise points times the square root of the sample size[13]. High enough confidence value implies an existence of a »signal« over and above noise and randomly distributed data.

Having the rough picture of cluster number and optimal point distance determined, DBSCAN algorithm was implemented on the input matrices containing the weather and dispatch cause group cardinality data through the series of linear algebra calculations. The calculations, as well as graphical representations of the results, were done in Matlab. Finding the clusters is always a recursive process starting at the random data point that satisfies the DBSCAN core point criteria (there are at least MinPts other points in its ε-environment). After the first cluster is formed following the point proximity criteria (at least one cluster's core point in ε-environment), analyses continues from the first core point outside the cluster, while there are core points left. As the result, all the points were either sorted into one of the formed clusters or marked as the noise points. Different clusters represents the different weather type, so the clusters that contain mostly days (data points) with high number of cases (based on the 75th percentile of the distribution of daily number of cases) were specially marked to draw the conclusions. Same was done for the clusters with that contain mostly the days with low case count (based on the 25th percentile of the distribution of daily number of cases).

Results

In order to illustrate the motivation of the optimal distance choice, distance graphs for two dispatch cause groups are given on Figure 3. Upper limit for the optimal distance represents the value where the graph makes a sharp curve up, which leaves the distant noise points out of the consideration to form a cluster. However, in order not to exclude any potential signal points from the analyses, optimal distance must not be set too low either and this lower limit is determined by the minimum signal to noise ratio, which defines the confidence level[7] and should not have the lesser value then 0.157. According to their distance graphs, the optimal distances for breathing difficulties and stroke are 1.75 and 1.50 Euclidian distance points respectively. In comparison to the breathing difficulties, stroke's data points are closer to each other which imply the lower optimal point distance. Complete overview of optimal distances to be used with DBSCAN, for all the considered groups of dispatch causes are presented in Table 1.

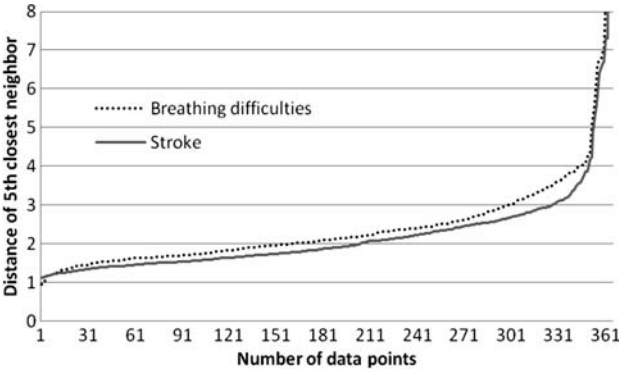

Fig. 3. 5th closest neighbor distance graph for two of the most frequent causes of dispatch. Values on x-axis mark the cardinality of data points (days) that have its 5th closest neighbor at least at the distance marked at y-axis or closer.

Results of the DBSCAN procedure are total of five vectors with 365 non-negative integers, representing the days of the year, each. The value of each vector member

TABLE 1
DBSCAN ALGORITHM PARAMETERS FOR THE MOST FREQUENT CAUSES OF DISPATCH

Dispatch cause group (MinPts=5)	ε
Breathing difficulties	1.80
Suspicion of stroke	1.50
Chest pain and suspicion of heart disease	1.75
Headache or dizziness	1.75
Abdominal or urinary tract	2.00

TABLE 2
DBSCAN ALGORITHM RESULTS FOR THE MOST FREQUENT CAUSES OF DISPATCH

Breathing difficulties	SNR=0.83	CONF=15.93	
Cluster number	TC	HCP (%)	LCP (%)
Cluster 1	75	4.00	13.33
Cluster 2	6	100.00	0.00
Cluster 3	72	5.56	16.67
Cluster 4	8	0.00	87.50
Cluster 5	5	0.00	0.00
Stroke	SNR=0.75	CONF=14.26	
Cluster number	TC	HCP (%)	LCP (%)
Cluster 1	7	0.00	100.00
Cluster 2	87	3.45	22.99
Cluster 3	28	3.57	28.57
Cluster 4	28	0.00	17.86
Cluster 5	6	0.00	0.00
Chest pain or heart disease	SNR=0.73	CONF=13.95	
Cluster number	TC	HCP (%)	LCP (%)
Cluster 1	110	8.18	13.64
Cluster 2	40	7.50	2.50
Cluster 3	5	0.00	100.00
Headache or dizziness	SNR=1.85	CONF=35.34	
Cluster number	TC	HCP (%)	LCP (%)
Cluster 1	228	16.67	10.53
Cluster 2	9	22.22	0.00
Abdominal or urinary tract	SNR=1.74	CONF=33.32	
Cluster number	TC	HCP (%)	LCP (%)
Cluster 1	225	6.22	20.89
Cluster 2	7	100.00	0.00

SNR – Signal to noise ratio defined as cluster to non-cluster points ratio, CONF – Confidence of the cluster analyses for the dispatch cause group defined as SNR times square root of the data sample (365 days in the year), TC – Total number of data points (days) in the cluster, HC – Percentage of high case count days in the cluster based on the 75th percentile of daily count distribution, LC – Percentage of low case count days in the cluster based on the 25th percentile of daily count distribution

indicates their cluster affinity, except for the zero valued ones which are considered as the noise points. By sorting the data points into different clusters, DBSCAN recognizes distinctive types of weather and dispatch cause frequencies that might be associated to them. Table 2 summarizes DBSCAN findings by presenting the cluster cardinality for all the dispatch cause groups taken into consideration. Additional two columns present the number of days with considerably high and considerably low number of reported cases in the cluster respectively.

Breathing difficulties

DBSCAN procedure revealed total of five distinctive point clusters in the data space. Day of the year graphical review of the cluster distribution (Figure 4) clearly indicates the season pattern that symptoms of breathing difficulties follow. Clusters number 1 and 2 contain mostly the days of the winter period (November to March), as cluster number 3 covers the milder climate towards the middle of the year. Spring and summer weather recorded during the days belonging to cluster number 4 and 5 reported extremely low numbers of cases. As to the number of cases, cluster number 2 defines the climate with high dispatch cause frequency, as it contains only the days with high dispatch cause count, while cluster number 4 coincidences with extremely low number of reported cases of breathing difficulties (7 of 8 its days have the low dispatch cause count). Weather that this cluster captures can be described as the spring weather with no wind, lower mean relative humidity (60 to 80%) and considerable differences between day and night temperatures.

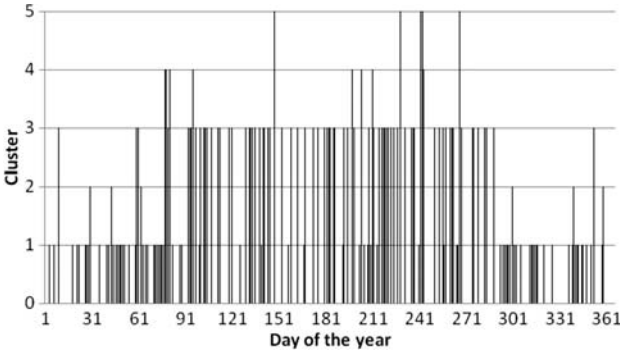

Fig. 4. Day of the year distribution of the cluster affinity for the breathing difficulties symptoms. The clusters make an apparent distinction between the warmer and colder seasons of the year.

Symptoms with suspicion of stroke

Five clusters were discovered among the points in suspected stroke symptoms data space as well (Figure 5). Cluster number 2 does not describe any particular season or weather as its data points are quite uniformly distributed over the whole year. Cluster number 3, 4 and 5 define the warmer spring and summer weather without containing a considerable number of days with low or high dispatch cause count, so there is no any association

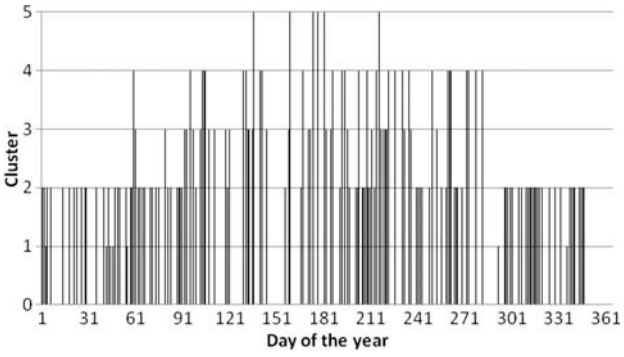

Fig. 5. Day of the year distribution of the cluster affinity for the suspected stroke symptoms. While days in cluster 2 are uniformly distributed throughout the year, other clusters clearly indicate seasonal changes.

between the weather and the dispatch cause frequency either. However, cluster number 1, although small in cardinality, groups only the cold, humid winter days of low count of cases.

Chest pain and suspicion to heart disease

Clusters that formed out of the chest pain or heart disease dispatch cause group were three distinctive sets of 110, 40 and 5 data points respectively. As the first and the second cluster capture both data points of high and low dispatch cause count they cannot be used to point out an association of any kind to the weather. All days grouped in third cluster tends though to describe the humid, autumn weather with the low count of this dispatch cause group.

Headache or dizziness

The analyses have not revealed any particular pattern that could associate some type of weather to the Headache or dizziness symptom group daily frequency. Although as much as 237 data points were classified as signal, huge majority went into the same cluster leaving only 9 in another. None of the two discovered data clusters contain a noticeable number of high or low case counts.

Abdominal or urinary tract symptoms

Two clusters were discovered in this data space, larger of which containing 225 of 232 clustered data points of diverse dispatch cause cardinality, randomly distributed throughout all seasons of the year. Second cluster put together only the days with high case count though without any distinctive weather marks except the high value of mean relative humidity.

Discussion and Conclusion

Obtained results indicate the existence of several associations between specific kind of weather and the occurrence of the days with high or low case count. Signal to noise ratio based method to determine confidence in the results of cluster analyses was designed primarily to find a small patterns in a huge data samples rather than obey the p value based methods of determining the uncertainty of outcome. However, there is a clear connection to these traditional concepts, as confidence defined on this way becomes greater as the p value becomes smaller[13,14].

Depending of dispatch cause group and its association with the weather conditions, DBSCAN algorithm made some apparent distinction between the different types of weather in different seasons of the year. This distinction was especially successful for breathing difficulties and suspected stroke symptoms where five different seasonal weather types were recognized. As to the breathing difficulties, although there are two distinctive weather clusters clearly associated with high and low case count respectively, it should be noted that there are many days of distinguishable case count that did not end up in those or in any other cluster. If successful, cluster analyses can roughly define a kind of weather and point out that it is associated with certain number of cases, or even the air pollution[15], but the association is not symmetric. Similar can be noted for the weather associated suspected stroke symptom counts, as well as for the rest of the findings.

Acknowledgements

Authors would like to point out their special gratitude to University hospital in Uppsala for accommodating the access to the emergency cases data, as well as to Uppsala University Department of Earth Sciences for providing the daily meteorological reports. Special thanks to Per Andersson (head of department) and Hans Blomberg (senior physician) from Uppsala University hospital's ambulance services department.

REFERENCES

1. VILLENEUVE PJ, LEECH J, BOURQUE D, Int J Biometeorol, 50 (2005) 48. DOI: 10.1007/s00484-005-0262-6. — 2. SZYSZKOWICZ M, Int J Occup Med Env, 20 (2007) 241. DOI: 10.2478/v10001-007-0024-2. — 3. ANDERSON HR, ATKINSON RW, BREMNER SA, MARSTON L, Eur Respir J, 21 (2003) 39. DOI: 10.1183/09031936.03.00402203. — 4. DOCKERY DW, POPE CA, Annu Rev Public Health, 15 (1994) 107. DOI: 10.1146/annurev. pu.15.050194.000543. — 5. MAKIE T, HARADA M, KINUKAWA N, TOYOSHIBA H, YAMANAKA T, NAKAMURA T, SAKAMOTO M, NOSE Y, Int J Biometeorol, 46 (2002) 38. DOI: 10.1007/s00484-001-0110-2. — 6. JONES SA, JOY MP, Health Care Management Science, 5 (2002) 297. DOI: 10.1023/A:1020390425029. — 7. SUN Y, HENG BH, SEOW YT, SEOW E, BMC Emergency Medicine, 9 (2009) 1. DOI: 10.1186/1471- 227X-9-1. — 8. ABDEL-AAL R, MANGOUD AM, Comput Meth Prog Bio, 56 (1998) 235. DOI: 10.1016/S0169-2607(98) 00032-7. — 9. BREUER HW, BREUER J, FISHBACH-BREUER BR, Eur Arch Psychiatr Neurol Sci, 235 (1986) 367. DOI: 10.1007/BF00381006. — 10. ESTER M, KRIEGEL HP, SANDER J, Xu X, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings (2nd International Conference on Knowledge Discovery and Data Mining, Portland, 1996). — 11. GUETING RH, The VLDB Journal,

3 (1994) 357. DOI: 10.1007/BF01231602. — 12. LUZ PM, MENDES BVM, CODEÇO CT, STRUCHINER CJ, GALVANI AP, Am J Trop Med Hyg, 79 (2008) 933. — 13. SACKETT DL, CMAJ, 165 (2001) 37. — 14. COOK RJ, SACKET DL, BMJ, 310 (1995) 452. DOI: 10.1136/bmj.310.6977.452. — 15. GVOZDIĆ V, BRANA J, MALATESTI N, PUNTARIĆ D, VIDOSAVLJEVIĆ D, ROLAND D, Coll Antropol, 35 (2011) 1135.

J. Malkić

OpenLink International GmbH, Friedrichstrasse 200, 10117 Berlin, Germany
e-mail: jasmin.malkic@etfbl.net

KLASTERSKA ANALIZA VREMENSKIH I DNEVNIH ČIMBENIKA POVEZANIH UZ BROJNOST HITNIH INTERVENCIJA

S A Ž E T A K

Cilj istraživanja bio je procijeniti povezanost različitih meteoroloških uvjeta i broja hitnih slučajeva za pet karakterističnih skupina simptoma prijavljenih Dispečerskom centru službe hitne pomoći u Uppsali, Švedska. Odgovornost službe proteže se na region površine 8,209 km^2 sa 17 ambulanti i oko 320,000 stanovnika, koji predstavljaju ciljnu pacijentsku skupinu. Izvor medicinskih podataka za ovu studiju je baza podataka hitnih slučajeva za 2009. godinu, dok su metrološki podaci preuzeti iz godišnjeg vremenskog izvještaja Odjela za geoznanosti Sveučilišta u Uppsali. Medicinski i meteorološki podaci su integrirani u jedinstven podatkovni prostor gdje svaka točka predstavlja jedan dan sa svojim vremenskim parametrima i brojnošću pojedine skupine slučajeva. Na pet takvih podatkovnih prostora provedeno je ujednačavanje varijance i ortogonalna linearna transformacija (PCA), a zatim je primijenjen DBSCAN algoritam za rudarenje podataka. Kao rezultat, otkriveno je nekoliko podatkovnih klastera koji ukazuju na određene vremenske uvjete, brojnost slučajeva, kao i povezanost ovih veličina. Najzanimljiviji nalaz je da posebna vrsta zimskog vremena formira klaster sastavljen od dana sa povišenim brojem simptoma poteškoća sa disanjem, dok je jedan od klastera koji sadrži dane sa ljetnim vremenskim uvjetima pokazao sličnu povezanost sa niskim brojem ovih simptoma. Razina pouzdanosti nalaza procjena je na temelju omjera signala i smetnji u promatranim podatkovnim prostorima.