Метод профилей для селекции признаков из временных рядов в задачах анализа данных

П.С. Сташевский, И.Н. Яковина
Новосибирский государственный технический университет, Новосибирск, Россия

Аннотация: **В работе рассматривается задача генерации информативных признаков на основе временных рядов, которые упрощают использование, повышают информативность и качество различных методов анализа данных. Предложенный метод формирования профилей временного ряда позволяет использовать преобразованные ряды значений признаков для решения различных задач анализа данных. Применение предложенного подхода рассматривается на примере задачи оценки влияния погодных факторов на инфекционную заболеваемость по данным города Екатеринбурга за 200 5-2010 гг.**

Ключевые слова: **селекция признаков, временной ряд, интеллектуальный анализ данных, кластеризация, погода и инфекционная заболеваемость**

ВВЕДЕНИЕ

В настоящее время одной из самых сложных задач, связанных с применением методов интеллектуального анализа данных, в частности методов машинного обучения, является создание новых признаков (feature engineering) на основе имеющихся данных. Применение таких техник в некоторых случаях позволяет улучшить качество получаемых результатов, повысить их информативность и прозрачность для исследователя, а в некоторых случаях и упростить генерируемые модели [1]. Во многих случаях этот процесс никаким образом не формализован и полностью зависит от исследователя, работающего с данными [2].

В особой степени это актуально для задач исследования временных рядов, которые могут содержать шумы и выбросы, иметь сложную природу процесса, зависеть с точки зрения поведения от большого количества факторов. В частности, такой задачей является оценка влияния погодных факторов (температуры, давления, относительной влажности атмосферного воздуха, точки росы, скорости ветра и др.) на динамику инфекционной заболеваемости для передающихся водным путем патологий, рассматриваемая в данной работе. Применение напрямую методов интеллектуального анализа данных в такой задаче не позволяет получить моделей высокой точности и наглядности.

В связи с этим цель данной работы - это разработка метода селекции (генерации) информативных признаков из временных рядов с последующим применением более сложных техник анализа данных. Для достижения данной цели были поставлены следующие задачи:

- проанализировать особенности исходных данных на примере рассматриваемой задачи оценки связи погодных факторов и инфекционной заболеваемости для передающихся водным путем патологий;
- предложить метод извлечения информативных признаков из исходных данных и на его основе описать технологию решения задачи анализа временных рядов;
- провести эксперименты с использованием предложенного метода для данных города Екатеринбурга за 2005-2010 гг.

1. ОПИСАНИЕ ИСХОДНЫХ ДАННЫХ. ПОСТАНОВКА ЗАДАЧИ

В работе использованы данные из базы *Climate. Water. Diseases. Infections* версии 1.0 (*CliWaDIn* 1.0), содержащей информацию по различным городам России о погоде, состоянии водных источников и инфекциях, передающихся водным путем, а также общедоступного сервера «Погода России» (http://meteo.infospace.ru/).

В качестве исследуемых параметров, представленных в виде временных рядов, в работе рассматриваются: динамика среднесуточной температуры (T), давления (P), относительной влажности атмосферного воздуха (H), инфекционной заболеваемости (I) для г. Екатеринбург за 2005-2010 гг. (*Рис.* 1). Для приведенных временных рядов можно отметить следующие особенности:

1) во-первых, ряды являются нестационарными с определенной цикличностью (годовой);
2) во-вторых, для них характерно наличие шумов, выбросов и пропущенных данных, которые заранее были сглажены и восстановлены на приведенных графиках;
3) в-третьих, визуальный анализ не позволяет выявить какой-либо зависимости между инфекционной заболеваемостью и погодными факторами.

Приведенные данные используются для задачи оценки влияния погодных факторов на динамику инфекционной заболеваемости, которую в общем случае можно представить как:

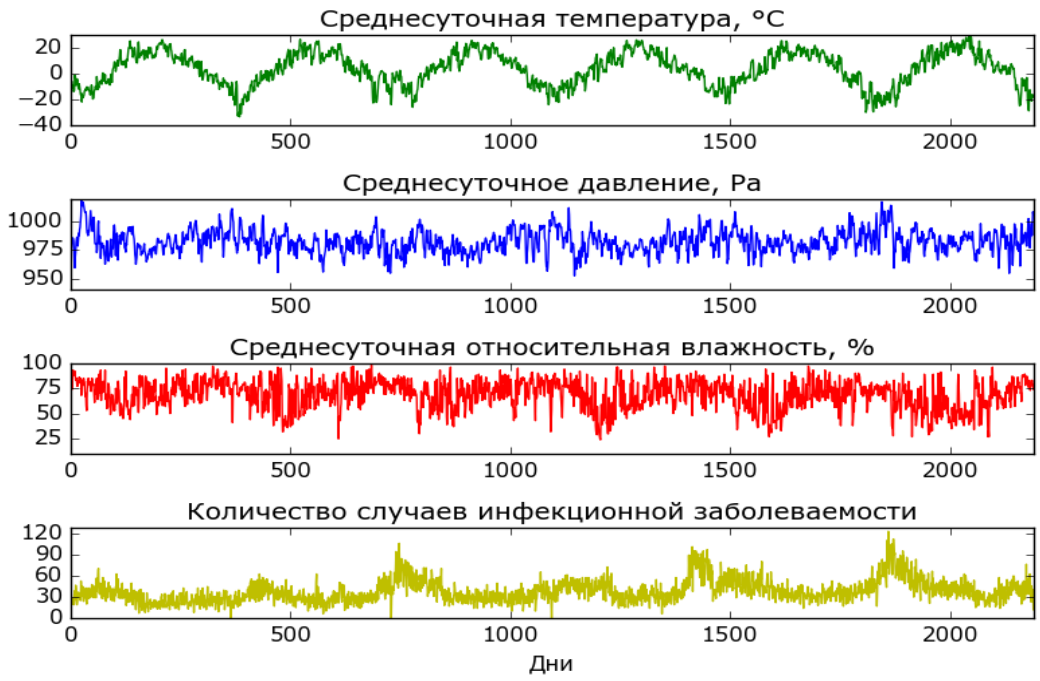

Рис. 1. Исходные данные (г. Екатеринбург, 2005-2010 г.)

$$\boldsymbol{M}(T,P,H) \rightarrow O(I), \quad (1)$$

где $\boldsymbol{M}$ – модель (или семейство моделей), позволяющих получить оценку O инфекционной заболеваемости I на основе значений погодных факторов температуры T, давления P, относительной влажности H.

Применение методов интеллектуального анализа данных, в частности машинного обучения, напрямую не работает в силу специфики рассматриваемой задачи. Инфекционная заболеваемость зависит от большого количества факторов, которые не могут быть учтены для построения моделей адекватной предсказательной точности и природа рассматриваемых факторов такова, что зависимости между ними носят сильно нелинейный характер, который усложняет стандартные подходы к анализу. В связи с этим необходима разработка метода, позволяющего создавать информативные признаки на основе имеющихся рядов, для их последующего анализа и построения оценочных моделей.

2. ТЕХНОЛОГИЯ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ ПРОФИЛЕЙ

Рассмотрим технологический процесс, используемый при решении задач анализа данных (*Рис*. 2). Как правило, он состоит из следующих этапов [3, 4]:

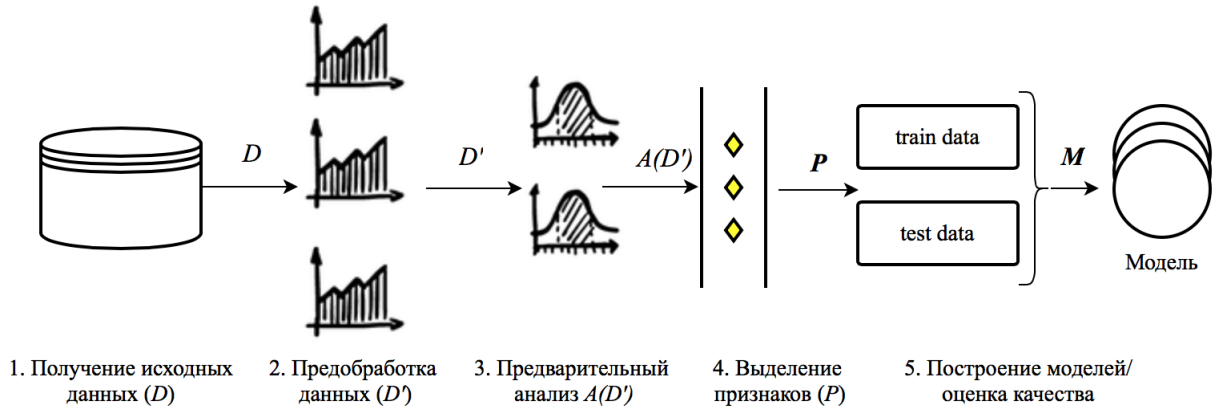

Рис. 2. Технологический процесс решения задач анализа данных

1) Извлечение исходного множества данных D из какого либо источника/источников данных (сетевые ресурсы, база данных и др.)
2) Предобработка данных ($D \rightarrow D'$) с целью восстановить пропущенные значения, удалить шумы и выбросы и т. д.
3) Предварительный (разведочный) анализ для поиска явно выраженных закономерностей, оценки распределений, статистических характеристик и качества исследуемых данных D'.

4) Выбор (селекция) признаков ***P***, которые получены с помощью каких-либо преобразованиями множества данных *D'*, либо взяты без изменений.
5) Применение к признакам ***P***, рассчитанным по исходному множеству данных *D'*, различных методов анализа данных для построения моделей и оценки их качества. Вид моделей зависит от класса решаемой задачи (восстановление регрессии, кластеризация, классификация и т. д.) и применяемого метода.

Четвертый этап в задачах анализа временных рядов предполагает, что получаемые новые признаки вычисляются на основе несложных вычислительных преобразований и легко интерпретируются с точки зрения решаемой задачи. Для этого в работе предлагается использовать профили. ***Профиль*** *временного ряда P* – это функционал, отображающий любые функциональные преобразования ряда в набор скалярных значений.

$$P(\Theta) = \Delta x, L(\Theta(x),t), \quad (2)$$

где $\Theta(x)$ – функциональное преобразование значения временного ряда, $L(\Theta(x),t)$ – агрегация полученных преобразованных значений по временным интервалам *t*.

В качестве примеров профилей можно привести следующие: вычисление недельных максимальных значений временного ряда, фильтрация значений ряда, которые ниже определенного порога с вычислением их средних суточных значений и др. Для рассматриваемых данных (*Рис*. 1) такими профилями могут быть: сумма по неделям значений температуры, которые находятся в определенном интервале, вычисление минимальных месячных значений, фильтрация значений заболеваемости, которые находятся выше определенного порога. Получение таких характеристик позволяет выявить тенденции в поведении временных рядов и успешно использовать их для построения моделей.

4. ОПИСАНИЕ ЭКСПЕРИМЕНТОВ

В качестве экспериментов в работе исследовалось решение задачи кластеризации погодных факторов и инфекционной заболеваемости [6,7] с использованием не преобразованных исходных данных (температура, давление, относительная влажность, инфекционная заболеваемость) и данных, полученных с помощью преобразования исходных временных рядов в набор профилей. Эксперименты были проведены для г. Екатеринбург (2005–2010 г.) с использованием следующей схемы:

1) Получение и предобработка исходных данных: вычисление среднесуточных характеристик для погодных факторов, посуточная агрегация случаев заболеваемости для инфекционных патологий (*Рис*. 1).
2) Вычисление следующих профилей для исследуемых данных:

Максимальный профиль Pmax – логическое преобразование (*L*) суточных (*t*) значений временного ряда в зависимости от превышения заданного порога *U*.

$$P_{\max}(x) = \begin{cases} 1, \Leftarrow x \geq U \\ 0, \Leftarrow x < U \end{cases}. \quad (3)$$

Минимальный профиль Pmin – логическое преобразование суточных значений ряда в зависимости от условия: является значение ряда меньше значения порога *U*.

$$P_{\min}(x) = \begin{cases} 1, \Leftarrow x \leq U \\ 0, \Leftarrow x > U \end{cases}. \quad (4)$$

Интервальный профиль Pint - логическое преобразование суточных значений ряда в зависимости от условия: попадает значение ряда в интервал $[U_1,U_2]$.

$$P_{\text{int}}(x) = \begin{cases} 1, \Leftarrow U_1 \leq x \leq U_2 \\ 0, \Leftarrow x < U_1 \;\&\; x > U_2 \end{cases}. \quad (5)$$

Вычисление значений порогов осуществлялось на основе значений персентилей, полученных по распределениям рассматриваемых факторов. Значения персентилей, а также виды вычисляемых профилей для разных факторов, приведены в *Табл*. 1

Табл. 1. **Характеристики профилей**

Фактор	*Pmax*	*Pmin*	*Pint*
Температура	U=90%	U=10%	U1=40%, U2=60%
Давление	U=90%	U=10%	U1=40%, U2=60%
Отн. Влажность	U=90%	U=10%	U1=40%, U2=60%
Заболеваемость	U=80%	-	-

Для инфекционной заболеваемости рассматривался только максимальный профиль, поскольку дни с высоким уровнем заболеваемости представляют для нас особый интерес в силу специфики рассматриваемой задачи.

3) Применение к исходным данным и к полученным профилям алгоритма кластеризации *K-means* [5] и сравнение полученных результатов.

5. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

На *Рис*. 1 приведены графики исходных данных, которые показывают, что в данных присутствует годовая периодичность, а с точки зрения инфекционной заболеваемости – наблюдаются периоды с большим количеством заболевших преимущественно в весенний и осенний периоды.

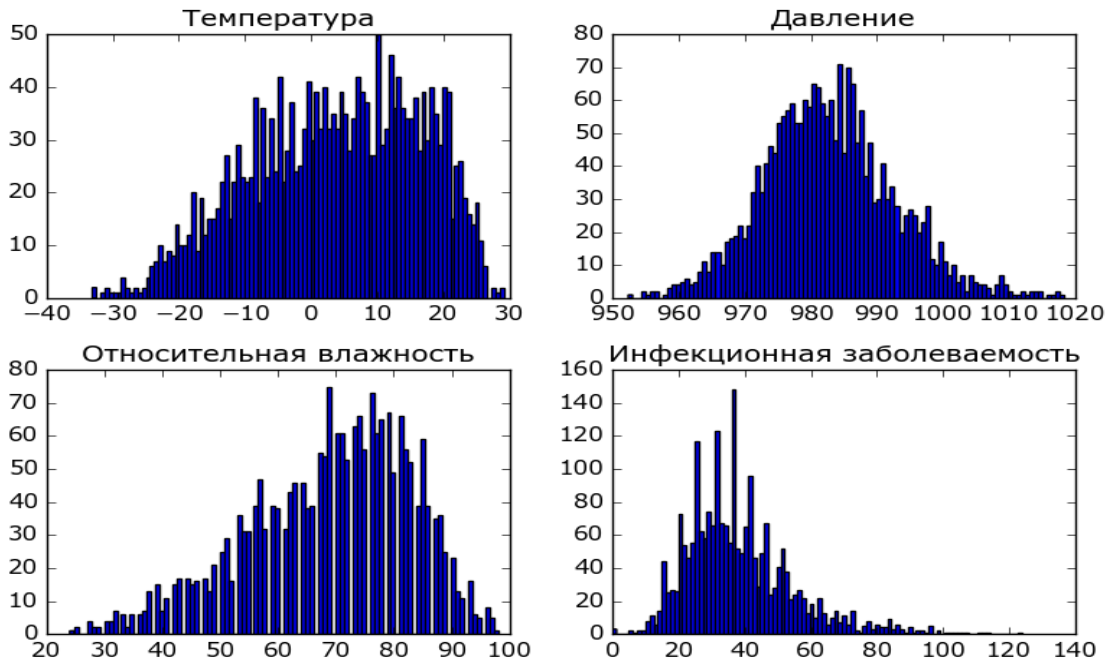

Рис. 3. Частотные распределения для исследуемых факторов

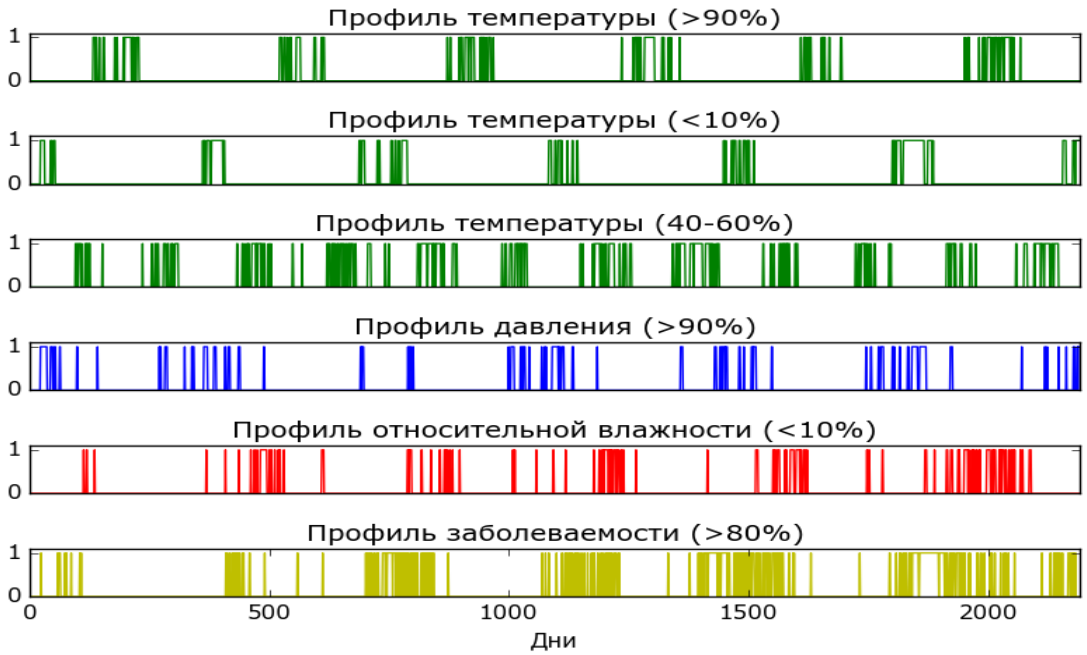

Рис. 4. Полученные профили температуры, давления, относительной влажности, инфекционной заболеваемости

Для исходных временных рядов были построены гистограммы, показывающие распределения исследуемых факторов (*Рис*. 3), на основе которых были рассчитаны значения персентилей, используемые в качестве пороговых значений для получения профилей.

На *Рис*. 4 приведены варианты профилей для погоды и заболеваемости. Полученные профили можно использовать не только в качестве новых информативных признаков далее в алгоритме кластеризации, а также в качестве метода разведочного анализа данных. Например, используя полученные профили, можно сделать выводы, что:

1) Построенные профили для температуры описывают смену сезонов и количество дней, которые можно отнести к одному или другому времени года. Это позволяет описать сезонные изменения (тренды), наблюдаемые при смене лет.

2) Профиль по максимальным значениям температуры практически не имеет совпадений с периодами высокой заболеваемости, в отличие от минимального и интервального профилей. Следовательно, можно предположить, что в жаркие период года фиксируется невысокое количество случаев инфекционной заболеваемости.

3) Визуальный анализ также показывает, что максимальный профиль давления и минимальный относительной влажности во многих точках совпадает с максимальным профилем заболеваемости. Следовательно, эти признаки также могу влиять на количество случаев инфекционной заболеваемости.

Полученные профили и исходные данные в необработанном виде были использованы в алгоритме K-means для получения 6 кластеров в каждом эксперименте. Сравнительные характеристики центров кластеров, полученные с использованием различных признаков, представлены в *Табл.* 2. Они показывают, что в эксперименте с использованием исходной выборки данных с точки зрения заболеваемости не произошло выделения кластеров с высокой заболеваемостью, что не позволяет оценить, при каких погодных условиях может произойти повышения количества случаев заболеваемости.

Табл. 2. **Результаты кластеризации**

№ кластера	Эксперимент 1 (исх. данные)				Эксперимент 2 (профили)					
	T, ^0C	*P, Pa*	*H, %*	*I*	*Pmax(T)*	*Pmin(T)*	*Pint(T)*	*Pmax(P)*	*Pmin(H)*	*Pmax(I)*
Кластер 1	11,60	994,74	57,05	14,91	0	0	1	0,05	0	0,10
Кластер 2	18,39	987,04	67,14	14,68	0	0	0	-	0	0,10
Кластер 3	- 15,16	1006,48	74,86	14,09	1	0	0	0	0	0
Кластер 4	-0,86	997,99	77,95	14,24	0	0	0,48	0,63	1	0,62
Кластер 5	- 28,32	1015,78	70,36	15,97	0	1	0	0	0	0,20
Кластер 6	-3,09	1004,99	57,14	16,07	0	0,38	0	0,67	0	0

Тогда как в случае использования профилей в качестве признаков произошло выделения кластера, в котором наиболее вероятно повышение инфекционной заболеваемости, со следующими характеристиками центра: низкая влажность атмосферного воздуха, возможно высокой давление и «межсезонье» с температурой, колеблющейся около 0 градусов.

ЗАКЛЮЧЕНИЕ

В работе предложен метод генерации информативных признаков для временных рядов, который в совокупности с методами интеллектуального анализа данных может повысить интерпретируемость результата, улучшить качество и упростить сложность получаемых моделей.

Данные метод был апробирован на задаче оценки влияния погодных факторов на инфекционную заболеваемость для г. Екатеринбург (2005-2010 г.), где с применением кластерного анализа был получены характеристики дней, для которых наиболее вероятны вспышки случаев инфекционной заболеваемости.

ЛИТЕРАТУРА

[1] Keogh, E., and Pazzani, M. 2000a. “A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases.” Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto,Japan.

[2] Leekha S, Diekema DJ, Perencevich EN. Seasonality of staphylococcal infections. Clin Microbiol Infect. 2012;18(10):927-33.

[3] McDonald LC, Banerjee SN, Jarvis WR. Seasonal variation of Acinetobacter infections: 1987-1996. Nosocomial Infections Surveillance System. Clin InfectDis. 1999;29(5):1133-7.

[4] Eamonn Keogh Michail Vlachos, Jessica Lin and Dimitrios Gunopulos. A wavelet-based anytime algorithm for k-means clustering of time series, 2003.

[5] Pierre Geurts. Pattern extraction for time series classification. LectureNotes in Computer Science, 2168:115–127, 2001.

[6] Сташевский П. С. Применение климатического профиля для исследования сезонной динамики климата / П. С. Сташевский, И. Н. Швайкова // Десятое сибирское совещание по климато-экологическому мониторингу : тез. Рос. конф., 14-17 окт. 2013 г. - Томск : Аграф-Пресс, 2013. - С. 145-146.

[7] Сташевский П. С. The study of climatic factors influence on the seasonal dynamics of infectious pathologies / П. С. Сташевский, И. Н. Яковина // Proceedings of IFOST2012. The 7th International Forum on Strategic Technology IFOST2012. Volume 1. – Tomsk Polytechnic University, September 17-21, 2012, p. 628-632.

Сташевский Павел Сергеевич – к.т.н., доцент кафедры вычислительной техники НГТУ. Автор более 20 научных работ. Сфера научных интересов: методы машинного обучения, интеллектуальные системы.
Email: stashpavel@gmail.com

Яковина Ирина Николаевна – к.т.н., доцент кафедры вычислительной техники НГТУ. Автор более 40 научных работ. Сфера научных интересов: методы интеллектуального анализа данных, интеллектуальные системы, робототехника. Email: irina.nir@gmail.com

Method of feature engineering for time series in data analysis problems

P.S. STASHEVSKIY, I.N. YAKOVINA

Abstract. The problem of generating informative features based on time series can make models easier to use and improve the the quality of various data analysis methods. We introduce technology which includes the step of calculating the time series profiles, then applying the received characters in higher-level machine learning techniques. We use this approach to solve the problem of communication of weather factors and infectious diseases. Experimental results obtained on the data for the city Yekaterinburg for 200 5-2010 years.

Key words: feature selection, time series, data mining, clustering, weather and infectious diseas

REFERENCES

[1] Keogh, E., and Pazzani, M. 2000a. “A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases.” Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto,Japan.

[2] Leekha S, Diekema DJ, Perencevich EN. Seasonality of staphylococcal infections. Clin Microbiol Infect. 2012;18(10):927-33.

[3] McDonald LC, Banerjee SN, Jarvis WR. Seasonal variation of Acinetobacter infections: 1987-1996. Nosocomial Infections Surveillance System. Clin InfectDis. 1999;29(5):1133-7.

[4] Eamonn Keogh Michail Vlachos, Jessica Lin and Dimitrios Gunopulos. A wavelet-based anytime algorithm for k-means clustering of time series, 2003.

[5] Pierre Geurts. Pattern extraction for time series classification. LectureNotes in Computer Science, 2168:115–127, 2001.

[6] Stashevskij P. S. Primenenie klimaticheskogo profilja dlja issledovanija sezonnoj dinamiki klimata / P. S. Stashevskij, I. N. Shvajkova // Desjatoe sibirskoe soveshhanie po klimato-jekologicheskomu monitoringu: tez. Ros. konf., 14-17 okt. 2013 g. - Tomsk : Agraf-Press, 2013. - S. 145-146.

[7] Stashevskiy P.S., Yakovina I.N. The study of climatic factors influence on the seasonal dynamics of infectious pathologies. Proceedings of IFOST2012. The 7th International Forum on Strategic Technology IFOST2012. Volume 1. – Tomsk Polytechnic University, September 17-21, 2012, p. 628-632.