NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Filipp Bykov, Vladimir Gordin

FORECASTING MOSCOW AMBULANCE TRIPS

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: SCIENCE, TECHNOLOGY AND INNOVATION

WP BRP 36/STI/2015

Filipp Bykov[1], *Vladimir Gordin*[2]

FORECASTING MOSCOW AMBULANCE TRIPS

This paper presents a method and computational technology for forecasting ambulance trips. We used statistical information about the number of the trips in 2009-2013, the meteorological archive, and the corresponding archive of the meteorological forecasts for the same period. We take into account social and meteorological predictors simultaneously. The method may be used operatively for planning in the ambulance service. It may be applied for all trips and for specific subgroups of diseases. The method and the technology may be applied for any megalopolis if the corresponding medical and meteorological information is available.

Keywords: weather forecasting, trips forecasting, disease, air temperature, correlation function, spline, optimization

JEL code: C32, C52, C53, C61, C63, I1

[1] Researcher, Hydrometeorological Research Centre of Russian Federation, E-mail: bphilipp@inbox.ru

[2] Professor, Department of Mathematics, Faculty of Economics, National Research University - Higher School of Economics and Leading Researcher, Hydrometeorological Research Centre of Russian Federation, E-mail: vagordin@mail.ru

Introduction

Ambulance Service in Moscow has a long history, see, e. g. http://www.mos03.ru/about/about.php. The operative data about the calls is available on this site.

The number of calls to the ambulance service in Moscow is about 5 million per year. About two thrids of the calls lead to ambulance trips. Here we analyse here only such calls.

These trips are not distributed uniformly. The number of the ambulance trips (NAT[3]) depends on the season, and the day of the week. The impact of the public holidays is also significant as is the meteorological situation. Until 2013 the list of medicine diseases contained more than 350 items (in 2014 the number tripled). The list includes fails callouts. The dynamics of NAT for various diseases varied significantly.

Since these statistics have some temporal correlation, i.e. information from previous days can improve (and improve significantly) the forecast of NAT in comparison to the "climatic" forecast, when we know mean values that obtained by appropriate averaging over an archive only.

We use the known statistics of NAT for some previous days (up to 35 days before) to forecast NAT for the following days (up to 28 days). We use this algorithm operatively, cyclically updating the available information and this provides a shift in the forecasting horizon.

The accuracy of such forecasts depends on their lead time, and on which disease the ambulance was called out for. For comparison we used the error of the inertial forecast (tomorrow there will be the same NAT as today or as week ago). Our method was twice as accurate as the inertial forecast.

We used the depersonalized database of trips during 2009-2013, that was kindly data provided by the A. S. Puchkov ambulance station and the meteorological databases of the Russian Hydrometeorological Centre. We evaluated the statistical regularities in the NAT overall and for separate groups of diseases, as functions of time, and to develop a method of forecasting these functions for various lead times. The problem of predicting NAT may be formulated in the different ways: we know or do not know NAT during the previous period.

[3] In (Sun 2014) the abbreviations EDV (emergency department visits) and EAD (emergency Ambulance dispatches) were used. In (Turner 2012) is referred as "number of ambulance attendances" and in (Murakami 2012) as "ambulance transports".

We obtained that NAT depends on the actual weather in the city. The weather's impact differs significantly for different diseases. We were interested in the accuracy of the forecast for the 12-hour sum of NAT in a real time, when the future weather is not known exactly. That is why we compare these results with the results of our weather forecasts (Bagrov 2014). We estimate the impact of the errors in these weather forecast on the forecast errors of the NAT.

The numerical model of weather forecasts for several days (both its method and results) was described in Bagrov (2014). We describe here the numerical forecasting using Moscow data only, but the described methods can be applied to any megalopolis, where there are the similar NAT statistics.

Our article is organised as follows. **Section 1** contains short review of recent publications about connection between weather and health. **Section 2** describes the statistics of the ambulance trips and their forecasting without concrete weather influence; **2.1** the separation of all trips into sub-groups according to the reasons for the callouts (diseases); **2.2, 2.3** the dynamics of the ambulance trips per week and per year. The impact of the statistics of the weekly period is explained by social factors only, unlike the yearly period, because there is an essential difference in the weather in Moscow between winter and summer. **2.4** lists several unexplained phenomena in the medical data. **2.5, 2.6** explain briefly (for details see **Appendix**) our computational approach to the statistical description of the NAT dynamics without the impact of the specific weather and its forecast. In **Section 3** we add meteorological predictors. **3.1** takes into account the real (or forecasted) weather. **3.2** describes the details of our computational experiments, comparing dependent and independent sampling. In **3.3** we consider the forecasting of NAT for cardiovascular diseases, because they correlate significantly with air temperature.

1. A short review of the association between weather and health

In the articles listed below the operative forecasting problem of the number of shifts (calls and hospitalizations) was not considered. However a general analysis of statistical information about impact of various factors on health may be useful for such forecasts.

The dependence of Q (NAT) in the region of Pudong (Shanghai, China) on air temperature and on the day of the week was considered (Sun 2014). The plots of dependence of Q on the mean diurnal air temperature are represented. Several versions of the approximation of the dependence in the interval of high temperatures were also given.

The dependence of Q on the maximal air temperature during heat in August 2010 in 47 Japanese prefectures (the mean value during the heat period of the maximal daily air temperature) was evaluated (Murakami 2012) as 1.8 shifts/$^{\circ}C$ per 100 000 habitants. A similar dependence of Q on air temperature was observed in Sydney, Australia in 2011 (Schaffer 2012), in Toronto, Canada in 2005 (Dolney 2006; Bassil 2011), and in Emilia-Romania, Italy, (Alessandrini 2011).

The dependence of the frequency of diseases (calls, hospitalizations, etc) for cardiovascular and respiratory diseases on various factors should be investigated separately. The dependence of NAT for these groups was considered in Makie (2002). The impact of the following factors was considered: air temperature (1, 3, or 7 days previously), the day of the week, and atmospheric pressure. The dependence on the air temperature was approximated by a piece-linear function.

The number of admissions to the hospitals in New York, USA in 1991–2004 (H) was analysed in Lin (2009). The dependence of H on air temperature in hot periods ($T_{air}^{day} \geq 27^{\circ}C$), and additionally on air humidity for any day of week j: $H \approx s\left(T-27\right)+b\cdot\left(T-27\right)_{+}+c_{j}$. The best results were obtained if the air temperature, which substituted in the formula, was obtained on the same day for the cardiovascular group, unlike the respiratory group, where a lag of 3 days is preferable. The impact of humidity is high if $T_{air}^{day} \geq 30^{\circ}C$.

The association of NAT in Brisbane, Australia during hot and cold weather in 2000–2008 was studied in Turner (2012). The optimal formulae for dependence includes data on air temperature, humidity with various lags.

Multiple regressions of hospital admissions in Shaghai, China for five types of coronary heart disease with air temperature and concentrations of NO_2, SO_2, PM_{10} was constructed in Xie (2012).

The frequencies of admissions for four types of diseases increase with any concentration of NO_2, SO_2, PM_{10} in the range 0,9–4,24%/(10μg/m^3). There is one exclusion (occult coronary heart disease), which grows weakly.

The dependence of the number of hospital admissions (H) with the diseases acute myocardial infarction in Melbourne, Australia in 1993–2004 was evaluated in Loughnan (2014). The number H was larger during temperature anomalies against climate values in the corresponding season. Growth of 1.5 times in 1994 (the anomaly was about 12^{O}C) and 2 times in 2004 (the anomaly was about

10^{O}C). We should establish that the dependence on the air temperature is not monotonic—there could be additional factors.

The arterial pressure of a patient is a natural measure of hypertension. The impact of weather on the health of such patients may be investigated quantitatively. The connection between air temperature and mean arterial pressure was approximated linearly in Chen (2013); the corresponding coefficient is estimate about $k \approx -0.25 \pm 0.05$mmHg/^{O}C.

The arterial pressure of patients may depend on physical activity, the time of day (the variation consists up to 30 mmHg), air temperature and the age of the patient (Goodwin 2001).

The air temperature dynamics for various groups of patients may influence the arterial pressure differently. The dependences were considered in Alpérovitch (2009) for men, women, smokers, nonsmokers etc.

Let us summarize the results. Health becomes worse as a result of adverse atmospheric conditions. In some articles quantitative assessments were obtained. Sometimes the results do not add up. To obtain non-trivial evaluations and results special singularities of groups of patients should be taken into account. Both social and meteorological factors are needed to predict NAT.

2. Forecasting NAT without the impact of weather

2.1. *Classification of diseases*

The number of possible diseases requiring an ambulance is very large, and we grouped them as follows:

I. Infectious diseases;

II. Diseases of the cardiovascular system;

III. Poisoning and diseases of the digestive system;

IV. Injury (including unsuccessful trips to injury);

V. Diseases of the nervous system;

VI. Respiratory viral infections;

VII. Diseases of the genitourinary system;

VIII. Fruitless trips.

The statistics and dynamics of NAT for diseases of these groups differ significantly. This article describes the 12-hour sum of trips for daytime (from 09:00 to 21:00 local time) and night (from 21:00 to 09:00 the following day).

We assume that the time t is measured in days and takes either integer values (for day) or half-integer values (for night). Let us define $Q_X(t)$ as NAT for disease X per 12 hours; $Q(t)$ is the total NAT for all diseases.

2.2. *Impact of the weekly tendency*

NAT (in total and by groups of diseases) significantly depended on the day of the week. The greatest NAT in most groups of diseases falls on a Monday (Tab. 1 up to the day of the week in bold). The maximal values for daytime and night are usually neighbours.

Weekly periodicity impacts the dynamics of NAT. We improve our forecasting if we preliminarily divide the time series into the typical values of the time series for the given day of the week.

Tab. 1. The average NAT $M_{Week}Q_X(t)$ and $M_{Week}Q(t)$ (the last column) per 12 hours by different groups of diseases on different days of the week, separately for daytime and night shifts

Group of diseases X	I	II	III	IV	V	VI	VII	VIII	∑
Night Sunday-Monday	**96**	**877**	**601**	533	**287**	**411**	**72**	148	**3025**
Monday	**192**	**1407**	**898**	956	**480**	770	**106**	293	**5102**
Night Monday-Tuesday	90	841	583	494	270	384	69	145	2876
Tuesday	181	1328	845	916	462	736	102	286	4856
Night Tuesday-Wednesday	89	823	563	482	268	376	67	142	2810
Wednesday	179	1323	838	906	468	730	100	288	4832
Night Wednesday-Thursday	87	810	556	484	267	372	68	142	2786
Thursday	176	1305	816	906	463	727	98	290	4781

Night Thursday-Friday	88	792	533	481	261	375	65	144	2739
Friday	180	1244	803	928	454	739	97	**300**	4745
Night Friday-Saturday	81	682	494	544	237	369	59	**152**	2618
Saturday	163	1134	746	**971**	426	734	88	279	4541
Night Saturday-Sunday	86	694	490	**575**	238	402	59	150	2694
Sunday	171	1221	794	969	438	**787**	93	264	4737
Mean for nights	88	789	546	513	261	384	66	146	2793
Mean for daytimes	177	1280	820	936	456	746	98	286	4799
Mean for days	133	1034	683	725	359	565	82	216	3797

2.3. *Long-term trend*

We give here some general information about NAT of the Moscow ambulance service. These dynamics total daily NAT and separately for daytime and night trips (Fig. 1a). We can see from the figures that public holidays and seasons influence the NAT. During the summer of 2010, the intense heat and smoke caused a sharp peak in NAT and there are strong oscillations during the New Year holidays.

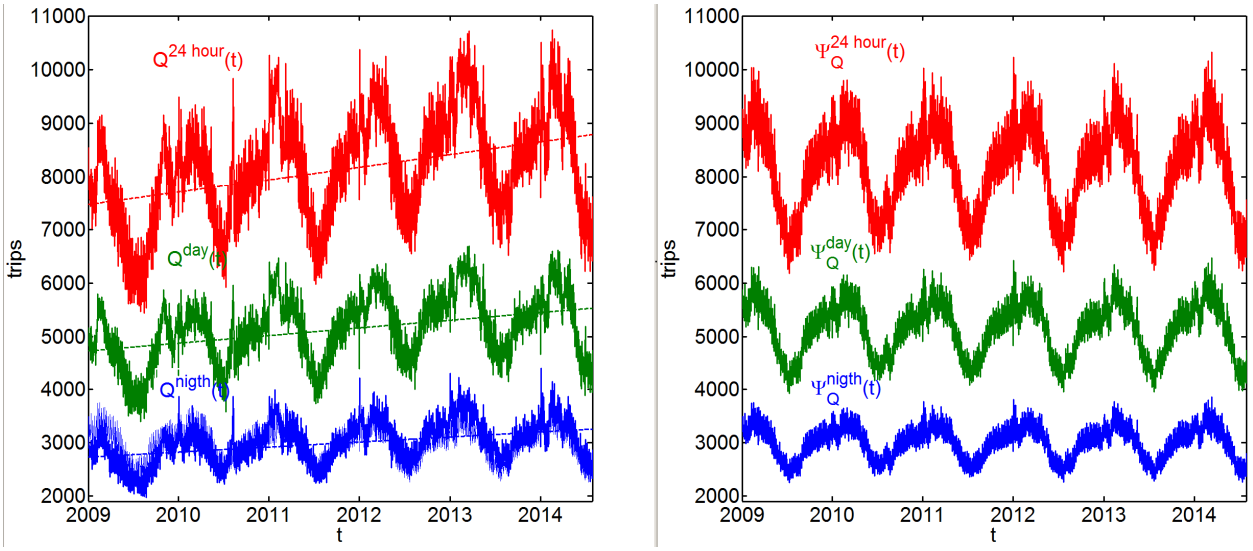

Fig. 1. a. The long-term trend changes NAT $Q(t)$. **b.** Calculated (see. Appendix) typical NAT $\Psi_Q(t)$ - the 28 years periodic function, which depends only on the time of year and day of the week. It also shows daytime, night shifts and their sum. On the horizontal axis are marked on January 1.

We introduce in the Appendix (by averaging per our data archive) for any group X a normalizing function $\Psi_X(t)$. The new function takes into account an impact of weekly oscillation in the function $\frac{M_{Week}Q_X(t)}{M_{Total}Q_X}$, that were described above, and the function $K_X(t)$ which is the dependence of the function on the specific day of the year (fig. 2, 3).

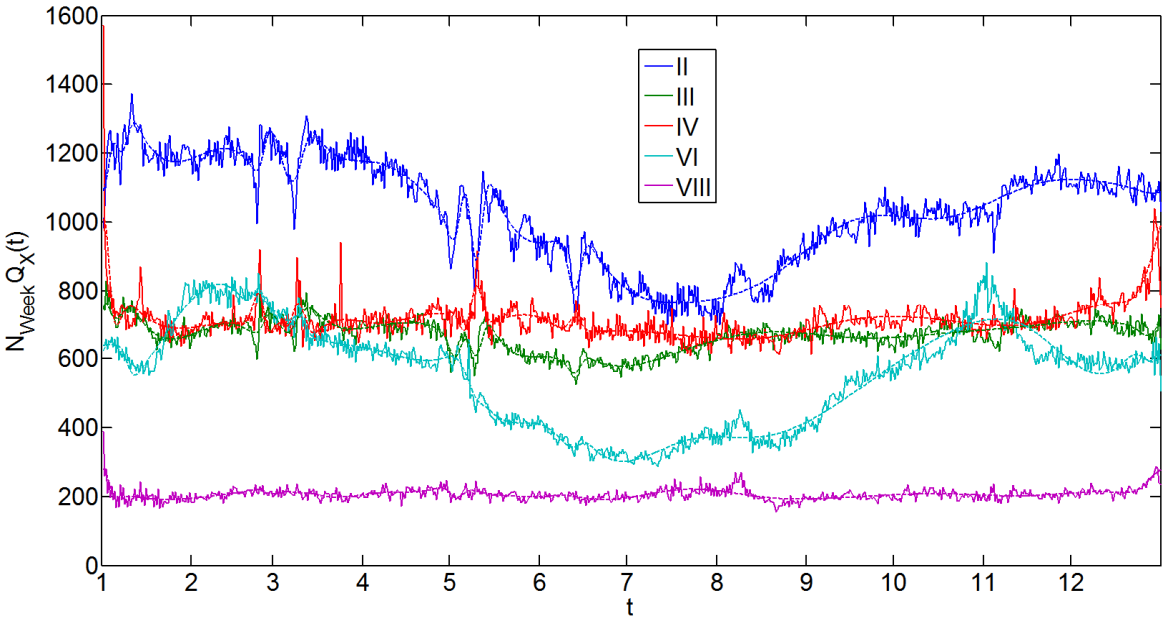

Fig.2. The mean NAT per half-days according to the archive of NAT for 2009-2013 (i. e., averaged over 5 years) for several groups of diseases. Another three groups do not have significant seasonal variations. In a leap year 2012 the data for February 29, were discarded. On the *t*-axis marked the first days of the months. The dotted graphs with the same colors (were used mean square splines) demonstrate the values after smoothing, see Appendix. Let us note the sharp peaks in the functions $M_{Year}Q(t)$ in public holidays: January 1, February 23, March 8, May 1 and 9, and June 12.

Since every forth year is a leap year, the period of the obtained function $\Psi_X(t)$ is equal to 28 years. Century amendments are ignored. We take into account the shifts of some weekends to join public holidays. See for details on such normalizing functions $\Psi_X(t)$ in the Appendix. We denote the normalizing function for all diseases as $\Psi_Q(t)$. Fig.1b presents normalizing functions $\Psi_Q(t)$ that are calculated for the functions $Q(t)$ presented in Fig.1a.

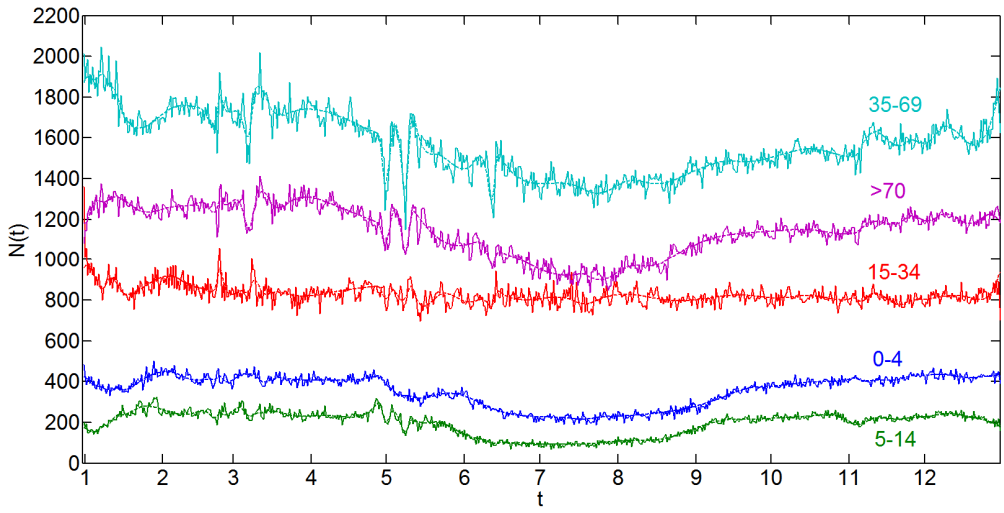

Fig.3. Same as on Fig. 2, but with separation for different age groups, instead of diseases. Note that the behavior of graphs for different age groups near holidays is significantly different. So, NAT decreases for 5-14 year olds and for 70+ year olds at New Year, and increases for other age groups. The weakest seasonal variations were observed for the age group of 15-34.

2.4. *Some unexplained observations of dynamics*

We cannot explain some phenomena in the dynamics of the NAT. Probably, they could be associated with administrative changes in the regulations for recording diseases. The first 3 diseases from the list below additively influence the dynamics of the total NAT: on Fig. 1 graphs $Q(t)$ for 2011-2013 on average are higher than $\Psi_Q(t)$ and in 2009-2010 they are below them.

1) Dorsopathy. To June 2010 (hereinafter inclusively) there are less than 50 trips per day and starting from 2011 more than 300. In the second half of 2010 we observed that growth of NAT is like a linear function of t.

2) Disorders of autonomic nervous system. Until July 2010 this disease is almost absent from the statistics. Afterwards, there is an increase with time in NAT per day, it is approximately proportional to the function $\sqrt{t}$, and by the end of 2013 this number is close to 220.

3) Chronic cerebrovascular disease. Until June 2010 it is about 25 trips per day, thereafter about 170.

4) Arthrosis, arthritis, polyarthritis. Until June 2010 it is equal to about 13 trips per day, afterwards it is equal to about 30.

5) Acute left ventricle failure. It is almost absent before June 2010. Then it is about 12 trips per day.

The plots for other diseases which do not have a sharp change, or correspond to values less than 5 trips per day.

2.5. *Normalized NAT*

We introduce instead of the function $Q_X(t)$, i.e. the NAT (total or for separate groups of diseases) the corresponding normalized function, i.e. the normalized NAT per day (or per night), which will be calculated by formula:

$$\Omega_X(t) = \frac{Q_X(t)}{\Psi_X(t)}. \qquad (1)$$

Such normalizations allow us to find unified regression coefficients together for different times of the year and days of the week, and then allow us to increase the volume of the archive when we construct the regression. We obtain regression coefficients with better precision. Additionally, the normalizations (1) lead to smoother correlation functions (CF) for the normalized processes (Fig. 4) than the CF of the original processes.

Since the normalizing function $\Psi_X(t)$ depends on the day of the year and of the half-day of the week only, it can be calculated initially, for any data. Therefore, if we can predict the normalized function $\Omega_X(t)$, then it is also possible to forecast the original function $Q_X(t)$; see (3).

These normalized functions of time $\Omega_X(t)$ can be considered as realizations of a stationary random process (Priestley 1981). Their correlation functions can be estimated by the standard formula

$$Corr_X(t) = \frac{M\left[\Omega_X(s)\Omega_X(s+t)\right] - \left\{M\left[\Omega_X(s)\right]\right\}^2}{\sigma^2\left[\Omega_X(t)\right]}.$$

Fig. 4 shows CF $Corr_X(t)$ for various groups of diseases *X*. An analysis of the plots shows that to aggregate the previously considered 8 groups of diseases into the following 3 larger super-groups, which although different from a medical point of view, have a similar CF:

(A) Diseases of the nervous system V and fruitless trips VIII;

(B) Viral respiratory infections VI;

(C) All other trips.

The correlation functions for different super-groups differ essentially.

Such aggregation into super-groups is useful if we are going to predict the total NAT. When we transform our method from 8 groups to 3 super-groups the accuracy of the prediction of *Q(t)* increases. It was confirmed by our numerical experiments.

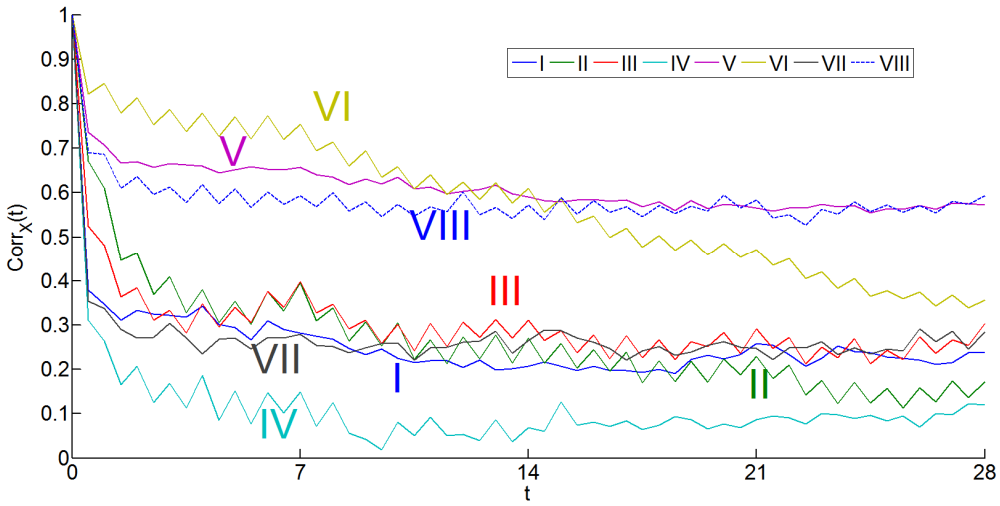

Fig. 4. The correlation function $Corr_X(t)$ for the normalized NAT for various groups *X* of diseases

3. Weather and forecasting algorithm

3.1. *Forecasting algorithm of the total NAT*

In this section we describe the forecasting algorithm for the total NAT *Q(t)* which uses dynamics of the NAT of the super-group *A, B, C* during the previous several weeks.

To forecast the total NAT *Q(t)* we predict the normalized NAT $\Omega_Q(t)$. We recorded unusual NAT for public holidays. Therefore we exclude them as predictors in the days that follow them. To avoid a distortion in our forecast after public holiday days, we form so-called "year without holidays"—the series $P_Q(t)$ instead of the series $\Omega_Q(t)$. The series $P_Q(t)$ distinguishes from $\Omega_Q(t)$,

but on holidays and after holiday days only. We use for such day t the value $\Omega_Q(t_*)$ in the last day $t_* < t$, which is not a holiday, Friday, Saturday or Sunday.

Let X be one of the super-groups of diseases (*A, B* or *C*). The average with weights (its decrease exponentially) for the previous 5 weeks by half-days NAT for a corresponding group of diseases are:

$$M_X(t,\alpha) = \frac{\sum_{k=0}^{14T-1} \exp\left(-\alpha \frac{k}{14}\right) P_X\left(t - \frac{k}{2}\right)}{\sum_{k=0}^{14T-1} \exp\left(-\alpha \frac{k}{14}\right)},$$

where the constant $\alpha > 0$ will be defined below.

The variation of the boundary of the averaging at the limits T=2–10 weeks dos not influence significantly on the final result of the forecasting. However, for the choice T=5 weeks we obtain the minimum forecast error (3).

According to Subsection 2.5 and Fig. 4, we aggregate our 8 groups into 3 super-groups of diseases that significantly differ from one another in their CF. So we obtain for the most numerous super-group C that the maximum impact for the forecast of 1–2 days is given by NAT yesterday and for the previous week. As for the forecasts with a longer lead time, the impact of NAT on the last day is not significant. The essential predictors for the super-group B (a group of respiratory viral diseases) give NAT for the last available day and the average NAT for the last available week. For the super-group A - we use as a principal predictor the mean weekly NAT only.

We use formula (2) for the forecasting the total NAT $\Omega_Q(t)$ (without any impact of the air temperature yet) with the lead time z days. Various super-groups impact the following formula differently:

$$\begin{aligned}\Omega_Q(t+z) \approx \Omega_Q^z(t+z) = a_1 + a_2 M_C(t,\alpha_C) + a_3 (P_C(t) - M_C(t,\alpha_C))(3-z)_+ \\ + a_4 M_B(t,\alpha_B) + a_5 M_A(t,\alpha_A), \qquad (2)\end{aligned}$$

where the weights $a_1,...,a_5$ and $\alpha_A, \alpha_B, \alpha_C$ need to be determined. Here the expression $(x)_+ = x$, when x>0, and =0 otherwise.

In order to choose the optimal weights, we minimized the mean square error of the forecast of the NAT, i. e. the function which takes into account the weight function Ψ_Q (see formula (1)) is:

$$\sum_{z,t}\left(\left(\Omega_Q(t+z)-\Omega_Q^z(t+z)\right)\Psi_Q(t+z)\right)^2 \to \min_{a_i,\alpha_X},$$

where the summation is over all half-days, i.e., z=1/2,1, 3/2...,28 and for all t, such that all values t+z in this formula lie inside the archive 2009-2013. The total weight function is obtained by summing the weight functions for all super-groups: $\Psi_Q(t)=\Psi_A(t)+\Psi_B(t)+\Psi_C(t)$.

The best values of the parameters $a_1,...,a_5$ were obtained by the method of least squares and the parameters $\alpha_A,\alpha_B,\alpha_C$ by the gradient descent method. The optimal values are represented in the upper row of the Tab. 2.

Tab. 2. The best values of numerical parameters for the forecasting by formula (2)

Impact of air temperature	α_A	α_B	α_C	a_1	a_2	a_3	a_4	a_5
No	0.59	2.72	0.74	0,648	-0,200	-0,578	0,085	0,844
Yes	0.64	2.83	0.62	0.435	-0,173	-0,283	0,073	0,776

The final forecast of the total NAT was calculated according to the formula

$$Q_z(t+z)=\Psi_Q(t+z)\Omega_Q^z(t+z). \qquad (3)$$

Fig. 5 represents the estimation of errors for various versions of the forecasting of NAT separately for daytime and night shifts. The plots of the errors in such division as functions of the variable z are proportional. The forecast error for night shifts is about 60% less than for day shifts — which corresponds to the real proportion of NAT (on day and night shifts) that are listed in the last column of tab. 3.

We can compute the normalizing function $\Psi_Q(t)$ preliminary for any time t. Therefore its lead time is “infinite”. The error of forecast (2) grows with lead time z relatively weakly (Fig. 5) and tends to errors of forecast $\Psi_Q(t)$. For the inertial forecast “today as week ago” the RMS error NAT per day is equal: for daytime shifts 269, 337, 370 for the forecast on the first week, the second and the third week respectively; for night shifts 196, 236, 258 respectively.

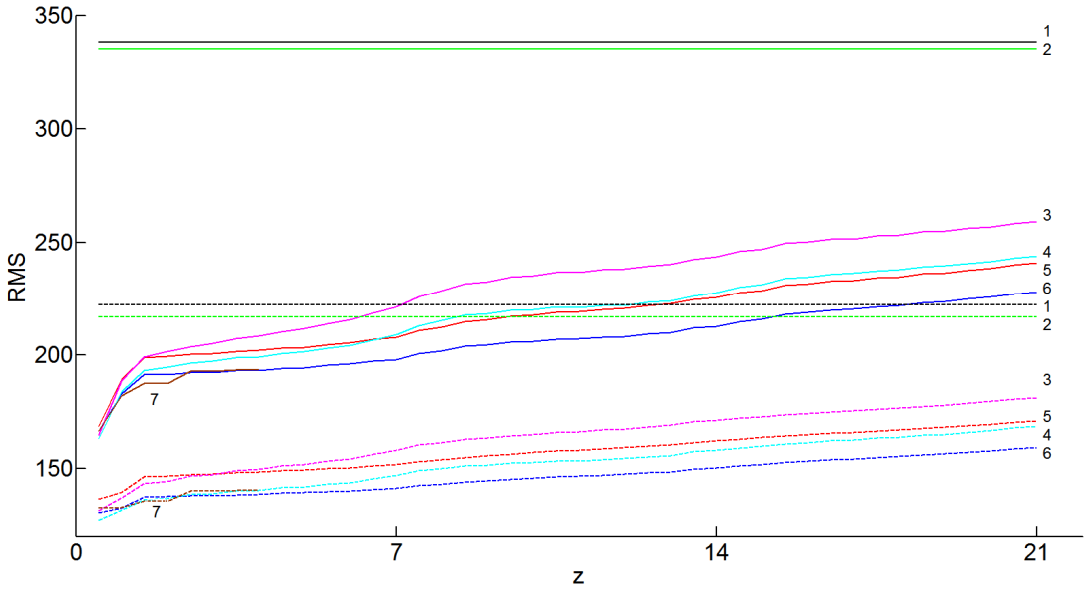

Fig. 5. The RMS error of the forecast of the total NAT per 12 hours depending on the lead time *z* (days). Data is divided into daytime shifts (solid line) and night (dashed line). 1 – the deviation $Q(t)$ from $\Psi_Q(t)$. 2 – the deviation $Q(t)$ from $\Psi_{Q,T_{\min},T_{\max}}(t)$. 3 – the error of the forecast, which uses NAT from several previous days, but without separation into super-groups of diseases; the data about air temperature were ignored. 4 - forecast without separation onto super-groups of diseases, but with impact of the temperature. 5 – we use the separation onto super-groups A, B, C, and do not use air temperature. 6 – we use the separation onto super-groups A, B, C, and take into account the air temperature for our forecasting. Here the air temperature was assumed to be known exactly (we use the information from the synoptical station). The real air temperature was exchanged on the forecasted one, according to (Bagrov 2014).

Fig. 5 represents the estimations of the errors of the forecasts, which take into account the real minimum and maximum daily air temperature in the city, see Subsection 3.2. The test sample used in Fig.5, 7 corresponds to the days on which we have forecasts of air temperature. The archive includes 1295 daytime and 1280 night shifts out of 1826 days (before March 2010 forecasts were not archived and small gaps exist due to computer failures). Constants were chosen for the full sample.

3.2. *Impact of the information on current weather*

There are seasonal tendencies in weather. Therefore, if we receive medical information about a specific city only, we need to separate the impacts of such predictors as “annual variations in the NAT” and “weather”. We need to determine the seasonal tendencies (e.g. associated with the

seasonal migration of the population) and the impact of the air temperature simultaneously, rather than sequentially.

Here we used the meteorological observation data from the synoptical station № 27605 Moscow-Baltschug, located in the centre of Moscow. We limited ourselves to studying the influence of daily maximal $T_{\max}$ and minimal $T_{\min}$ air temperature (during the corresponding full day) on the NAT. If we are talking about the assessment of daytime NAT, then $T_{\min}(t)$ is the half-sum of the minimal temperatures for the previous and the next nights for the day t. If we predict the night trips, then we use the half-sum of the maximum temperatures $T_{\max}(t)$ for the previous and next day.

Let us introduce the following more sophisticated normalizing function $\Psi_Q(t, T_{\min}, T_{\max})$. The corresponding definition is given in Appendix. We take into account again: what day of the year and of the week correspond to *t* and what it is daytime or night shift. Also the normalizing function depends now on daily maximal $T_{\max}$ and minimal $T_{\min}$ air temperature during the day.

Then we use the function

$$\Omega_{Q,T_{\min},T_{\max}}(t) = \frac{Q(t)}{\Psi_Q(t, T_{\min}, T_{\max})}.$$

Then we compute, as in Subsection 3.1, but with using the function $\Omega_{Q,T_{\min},T_{\max}}(t)$ instead of $\Omega_Q(t)$, and determine anew the constants a_i, $i = 1,...,5$ and $\alpha_X, X = A, B, C$ in formula (2) (see tab.2).

The RMS errors for several versions of the forecasts, depending on the lead time *z*, are represented in Fig. 5 (curves 3–7).

We can forecast the air temperature near the Earth's surface for a lead time of up to five days (Bagrov 2014). The forecast errors for NAT increase if we use the forecasted air temperature instead of the actual one. We can see in Fig. 5 the comparison of the error for the forecast of NAT when we used the forecasted temperature according to Bagrov (2014), and when we used the real air temperature.

We used for the forecast temperature the same parameters which were determined for the actual temperature. In final version of the forecast for NAT, it is preferable to use the parameters which are optimized according to the specific meteorological forecasting scheme.

3.3. *Organization of computing—dependent and independent sampling*

To estimate the quality of our forecasting schemes both for the total NAT and for separate groups of diseases we need to use the same available database that we used for the development and optimization of these schemes. For example, the parameters $a_1,...,a_5$ in formula (2) were chosen so that our sample RMS error for all dates was minimal.

If the sample is larger and therefore more representative, we can estimate more reliably the optimal values of the numerical parameters of our forecasting scheme. On the other hand, if we use all available data for such a choice, then there is the danger that the choice of the values is *ad hoc*, when these values are oriented exactly for this specific archive. In this case, the further application of such numerical parameters for new data will lead to a significant increase of the error compared to the error for the archive which we used for our parameters choice.

We tested our algorithm as follows: regression constants were chosen for the archive of 4 years (from the 5 available), followed by checking the accuracy of the forecast independently with the fifth year. This approximation error of the original function *Q(t)* gives a larger error than in the reference case (when the training set includes the archive for all 5 years). In 2009, the difference proved to be very prominent, compared to the rest of the years, see Tab. 3.

We give estimates only when the training set includes the archive for all 5 years. For the forecast errors with a lead time $z \le 21$ such estimates are shown in Fig. 5 (curves 6) separately for daytime and night shifts.

Tab. 3. The RMS of forecast NAT with lead time z = 1, 2, 3 days if for selecting optimal parameters used archive for all 5 years or only 4 (the corresponding year is not used).

Year	*z*	2009	2010	2011	2012	2013
The parameters choice by 5 years	1	144,9	163,4	154,2	161,9	144,5
	2	152,8	175,5	161,8	167,2	149,2
	3	154,9	177,1	164,4	167,5	149,1
The parameters choice by 4 years	1	182,1	167,3	165,3	168,5	153,0
	2	188,9	180,4	178,3	176,2	160,1
	3	192,8	182,4	183,0	176,8	160,3

3.4. *About forecasting NAT for cardiovascular diseases*

We can apply our methodology to forecasting NAT for any separate group of diseases if it is sufficiently large. Let us consider, as an example, the cardiovascular group II.

We divide it into two subgroups: G — hypertonia, and H — all other cardiovascular diseases. For the dynamics of NAT for these subgroups see Fig. 6. Cold weather contributes to the growth of NAT for G because blood vessels constrict.

The essential influence of T_{air} on arterial pressure was statistically confirmed in Chen (2013) on an archive which consisted of histories of 1831 patients in Shanghai, China, over three years.

We evaluate (see Fig.6) NAT with weekly correction $N_{Week}Q_X(t) = Q_X(t)\dfrac{M_{Total}Q_X}{M_{Week}Q_X(t)}$ the dependences for diseases subgroups $X = G,\ H$ on the time t (day of the year, day of the week, shift) $\Psi_X(t)$ and on air temperature $F_X(t) = f_X\left(T_{\min}(t), T_{\max}(t)\right)$. The sum of the functions $\Omega_{X,T_{\min},T_{\max}}(t)$ approximates NAT $Q_X(t)$ at time t. Both functions ($\Psi_X(t)$ and $\Omega_{X,T_{\min},T_{\max}}(t)$) are defined up to the additive constant.

Our numerical experiments confirmed that the impact of the information on air temperature $\Omega_{X,T_{\min},T_{\max}}(t)$ for the evaluation $Q_H(t)$ (for the subgroup *H)* is small (see Fig.6a). In other words, if we determined $\Psi_H(t)$, the impact of the real temperature T_{air} is small.

Conversely, the impact of the information on air temperature in the function $\Psi_{G,T_{\min},T_{\max}}(t)$ is more significant than impact of the seasons of year. But we need to consider the time of year: the impact of public holidays (see Fig.6b) cannot be evaluated by air temperature T_{air} only.

The forecast $\Psi_X(t)$ is possible for any lead time and its error is 15–25% smaller than the errors of any inertial forecast ("today is like yesterday" or "today is like week ago").

We approximate the dependence of NAT for subgroup G on the air temperatures by the formula:

$$f_G\left(T_{\max}, T_{\min}\right) = -0.096\left(T_{\max} + T_{\min}\right)^2 - 4.8\left(T_{\max} + T_{\min}\right) + 1264.$$

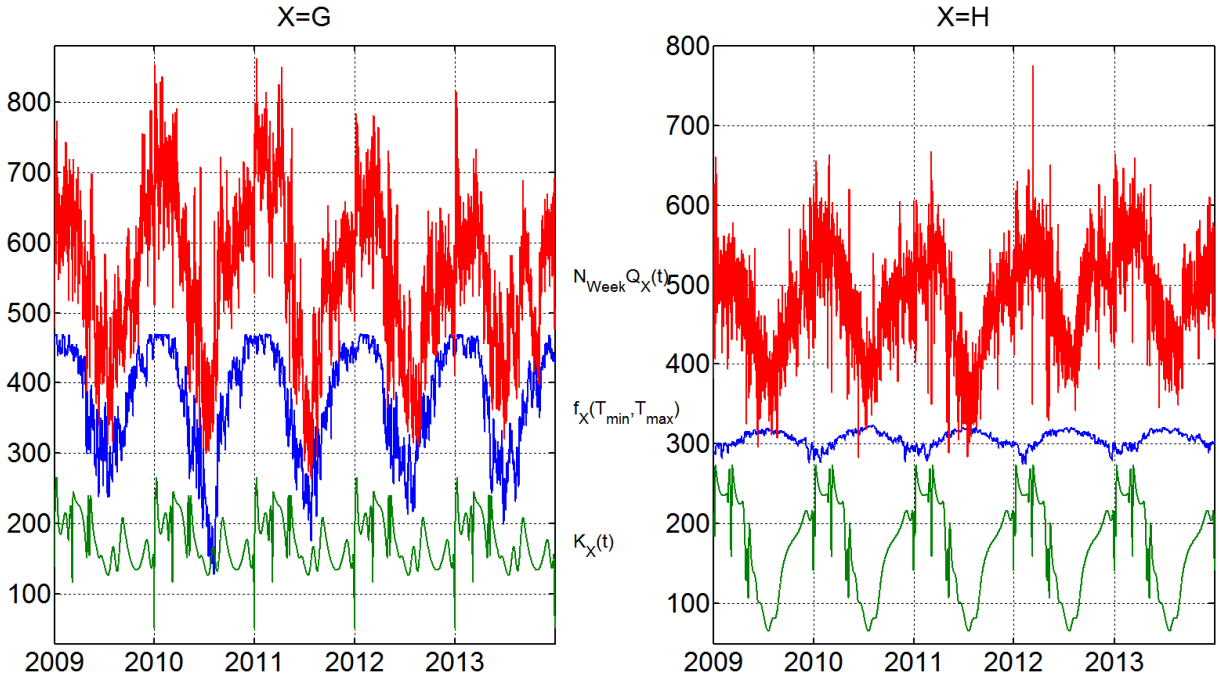

Fig. 6. The dynamics for the five and a half-year (2009-2014) period of time of NAT for cardiovascular group II. The NAT with weekly correction (red curve $N_{Week}Q_X(t)$); two graphs, which were determined by simultaneous minimization of the functional (4): averaged annual dynamics – smoothing spline (green curve) and the function $F_{II}(t) = f(T_{\min}(t), T_{\max}(t))$ - dependence of NAT on the maximal temperature (blue curve). The sum of the terms approximates the first curve. These two functions can be determined only with an accuracy of up to additive constant. 1-st January of the corresponding years is marked on the *t*-axis. We can see on the graph clearly visible peaks, near public holidays: January 1, February 23, March 8, May 1 and 9

We did not include such dependence for evaluation of $Q_H(t)$, since the improvement here was not significant, and we assumed $f_H \equiv 0$.

The results of the forecasting of NAT for the diseases of the cardiovascular group are shown in Fig.11. Here we used as predictors for the forecast only NAT of the cardiovascular group in the previous period. Information about the trips with other disease groups was ignored. If we add additional predictors to this regression, which take into account NAT for such groups that are not included in the group II, we can improve the result.

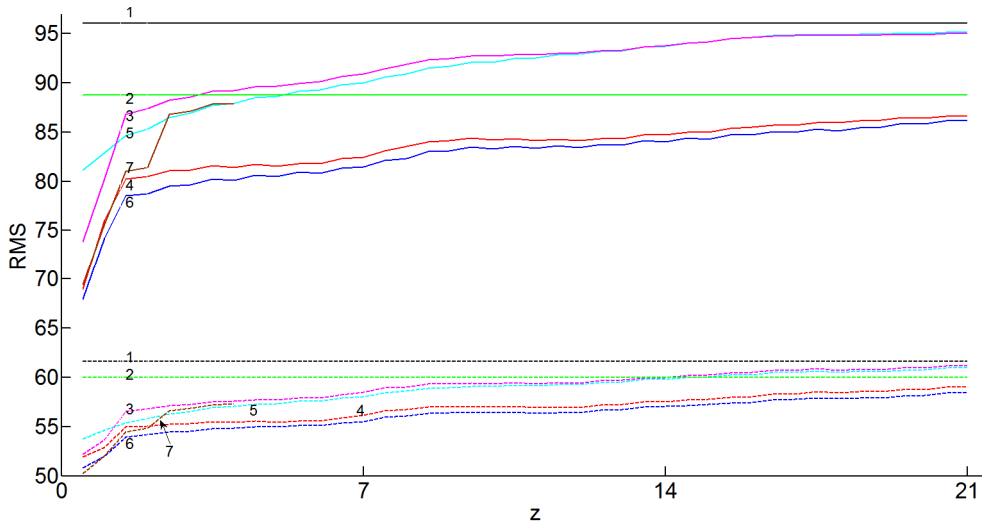

Fig. 8. The mean RMS error of the forecast of cardiovascular NAT per 12 hours depending on the lead time *z* (days). The dotted lines correspond to the independent sample (2013), solid lines – to the dependent one (2009-2012). 1 - deviation from the average calls value $\Psi_{II}(t)$; 2 - deviation from average values $\Psi_{II,T_{\min},T_{\max}}(t)$, which based on the real data about the air temperature; 3 - forecast without impact of the air temperature; 4 - forecast with impact of the air temperature. Here the temperature was assumed to be known exactly for the curves 2, 4, and 6. Curves 7 describe the forecast which is similar to 6, but it use the forecasted air temperature with corresponding lead time (see Bagrov 2014) instead of real air temperature.

We also estimated RMS error for the inertial forecast for the subgroups *G* and *H* in Tab.4. For subgroup *H* the error of the forecast "today is like yesterday" is larger than of the forecast "today is like week ago", and vice versa for the subgroup *G*. This can be explained by the significant weekly tendency for the NAT of subgroup *H*. The weekly tendency of the subgroup *G* is not significant.

In a separate consideration of trips during daytime and night shifts the graphic dependence on the lead time *z* of the forecast errors is proportional. Curves 7 on fig. 7 shows RMS error when we use the forecast air temperature.

Tab. 4. The mean RMS error for some forecasts of cardiovascular subgroup *G* and *H* NAT per 12 hours. The optimal version of the forecast written by bold font

Forecast method	$X = G$		$X = H$		$X = II$	
	Daytime shifts	Night shifts	Daytime shifts	Night shifts	Daytime shifts	Night shifts
Constant – mean value $M_{Total}Q_X$	148,6	82,2	96,7	63,9	221,7	134,8
Forecast depends only on the day of the week $M_{Week}Q_X(t)$	144,8	76,6	79,7	50,6	206,1	116,3
Today as yesterday: $Q_X(t-1)$	70,6	54,8	85,8	60,8	130,2	105,3
Today as week ago: $Q_X(t-7)$	97,9	57,5	62,1	43,6	127,0	82,2
Today as 2 weeks ago: $Q_X(t-14)$	109,9	62,9	69,2	47,0	145,5	90,5
Forecast according to the normalization function, depending on day of week and day of the year $\Psi_X(t)$	81,4	48,0	**48,1**	**34,1**	97,7	63,9
Forecast according day of week and air temperature (but no according day of the year)	81,1	46,4	58,0	40,1	110,0	73,2
Forecast according to the normalization function, depending on day of week and day of the year and air temperature $\Psi_{X,T_{\min},T_{\max}}(t)$	**69,9**	**43,2**	**48,0**	**34,0**	**89,0**	**62,1**

4. Conclusions and Discussion

This article describes the method of forecasting of number of ambulance trips (NAT) for various lead times. The errors of such forecasts are shown in Figs. 5. They 1.5–2 times less than the error of the inertial forecast of NAT. We forecasted separately NAT with diseases of the

cardiovascular group, see Fig.7, where it was demonstrated in particular the advantage of the forecasting method, when we use as a predictor the maximal daily air temperature. We used the medicine database for the Moscow. However, the developed method can be applied for other megalopolises and groups of diseases, if we have the appropriate medical and meteorological databases, but only if this group of diseases is not less several tens of cases per day.

Meteorological forecasts of surface air temperature according to the model (Bagrov 2014) introduce a relatively small additional error in the comparison to usage of the real meteorological data into the forecasting on a short lead time (several days) of NAT. The forecasted air temperature with a lead time that is more than 5 days seems unreliable.

We plan in the future to study the influence of other meteorological factors as well as chemical ones on the dynamics of the disease and, consequently, to predict the dynamics of efforts that are required from medical institutions.

We hope that the statistical evaluation of the influence of meteorological factors on the dynamics of medical problems may in some cases be useful for understanding the physiology of the disease and possible treatment options.

We can assimilate individual medical histories of those with specific diseases, and the relative meteorological archive. As a result we hope to evaluate how weather can influence the intensity of the disease. Knowledge of the weather forecast for several days will help us to predict states of health. A person can be proactive to avoid the anticipated worsening of their health.

We sincerely thank A. V. Sigachev for useful discussions.

The work was supported by a grant of the Economic Faculty of the National Research University - Higher school of Economics.

References

Alessandrini E, Zauli Sajani S, Scotto F, et al. *Emergency ambulance dispatches and apparent temperature: a time series analysis in Emilia-Romagna, Italy*. Environ Res 2011; 111:1192–200

Alpérovitch A, Lacombe JM, Hanon O, Dartigues JF, Ritchie K, Ducimetière P, Tzourio C. *Relationship between blood pressure and outdoor temperature in a large sample of elderly individuals: the Three-City study*. Arch Intern Med. 2009; 169(1):75-80. doi: 10.1001/archinternmed.2008.512

Bagrov AN, Bykov PL, Gordin VA. Complex forecast of surface meteorological parameters. Meteorology and Hydrology, 2014, No. 5, pp. 5-16 (Russian), 283-291 (English)

Bassil KL, Cole DC, Moineddin R, et al. *The relationship between temperature and ambulance response calls for heat-related illness in Toronto, Ontario, 2005*. J Epidemiology Community Health 2011;65:829–31

Bellman R, Kashef B, Vasudevan R. Mean square spline approximation. Journal of Mathematical Analysis and Applications, 1974, vol. 45, issue 1, pp. 47-53

Chen Q, Wang J, Tian J, Tang X, Yu C, Marshall RJ, Chen D, Cao W, Zhan S, Lv J, Lee L, Hu Y. *Association between ambient temperature and blood pressure and blood pressure regulators: 1831 hypertensive patients followed up for three years*. PLoS One. 2013; 8(12):e84522. doi: 10.1371/journal.pone.0084522

Dolney TJ, Sheridan SC. *The relationship between extreme heat and ambulance response calls for the city of Toronto, Ontario, Canada*. Environ Res 2006; 101:94–103

Goodwin J, Pearce VR, Taylor RS, Read KL, Powers SJ. *Seasonal cold and circadian changes in blood pressure and physical activity in young and elderly people*. Age Ageing. 2001 Jul;30(4):311-7

Lin S, Luo M, Walker RJ, et al. *Extreme high temperatures and hospital admissions for respiratory and cardiovascular diseases*. Epidemiology, 2009; 20:738–46

Loughnan M, Tapper N, Loughnan T. *The impact of "unseasonably" warm spring temperatures on acute myocardial infarction hospital admissions in Melbourne, Australia: a city with a temperate climate*. J Environ Public Health, 2014:483785. doi: 10.1155/2014/483785

Makie T, Harada M, Kinukawa N, Toyoshiba H, Yamanaka T, Nakamura T, Sakamoto M, Nose Y *Association of meteorological and day-of-the-week factors with emergency hospital admissions in Fukuoka, Japan*. Int J Biometeorol. 2002; 46(1):38-41

Murakami S, Miyatake N, Sakano N. *Changes in air temperature and its relation to ambulance transports due to heat stroke in all 47 prefectures of Japan.* J Prev Med Public Health, 2012; 45(5):309-15 doi: 10.3961/jpmph.2012.45.5.309

Priestley MB. Spectral Analysis and Time Series. Academic Press, 1981

Schaffer A, Muscatello D, Broome R, Corbett S, Smith W. *Emergency department visits, ambulance calls, and mortality associated with an exceptional heat wave in Sydney, Australia, 2011: a time-series analysis*. Environ Health, 2012; 11(1):3 doi: 10.1186/1476-069X-11-3

Sun X, Sun Q, Yang M, Zhou X, Li X, Yu A, Geng F, Guo Y. *Effects of temperature and heat waves on emergency department visits and emergency ambulance dispatches in Pudong New Area, China: a time series analysis.* Environ Health, 2014; 13(1):76 doi: 10.1186/1476-069X-13-76

Turner LR, Connell D, Tong S. *Exposure to hot and cold temperatures and ambulance attendances in Brisbane, Australia: a time-series study*. BMJ Open, 2012; 2(4) doi: 10.1136/bmjopen-2012-001074

Xie J, He M, Zhu W. *Acute effects of outdoor air pollution on emergency department visits due to five clinical subtypes of coronary heart diseases in Shanghai, China*. J Epidemiol. 2014;24(6):452-9

Appendix

NAT depends essentially on the time of the year. External conditions can change, e.g. average temperature, average rainfall and cloudiness. In addition, there are seasonal population movements (summer trips to dachas or vacations), which also affect NAT. The seasonal effect distinguishes different groups of diseases as well as different age groups, see Fig. 2, 3.

We need smoothed graphics for each group of diseases. In such smooth graphics short-period noises are filtered. These smoothed functions are shown in Fig. 2, 3 by the dotted lines of the same colour.

Let

$$M_{Total}Q_X = \frac{1}{N}\sum_{t=1}^{N} Q_X(t)$$

- be the total sample mean for an arbitrary numerical series $Q_X(t)$;

$$M_{Year}X(t)=\frac{1}{5}\sum_{k=0}^{4}X(t+365k),$$

- be a periodical function with a period of 1 year sample mean over 5 years (where t=1/2, 1, 3/2, …, 365);

$$M_{Week}Q_X(t)=\frac{1}{W}\sum_{k=0}^{W-1}Q_X(t+7k),$$

- be a periodical function with a period of 1 week (instead of a year), obtained by the averaging over all weeks—the sample mean over 5 years (where t=1/2, 1, 3/2,…, 7), W is the number of the weeks in the archive. The values $M_{Week}Q_X(t)$ are represented in Tab. 1.

All these averaged values are calculated separately for day and night shifts; t can be an integer or half-integer value, and the k only an integer.

We normalize the original time series which consists of NAT $Q_X(t)$ from its mean weekly values (see tab. 1). For this goal we define the normalization facto—the periodical function $F_{Week}Q_X(t)=\frac{M_{Total}Q_X}{M_{Week}Q_X(t)}$. The values of the multiplier are calculated by Tab. 1 or similar, and then we determine the normalized (non-periodical) function $N_{Week}Q_X(t)=Q_X(t)\bullet F_{Week}Q_X(t)$.

We calculate these functions $M_{Year}\left[N_{Week}Q_X(t)\right]$ and $M_{Year}\left[F_{Week}Q_X(t)\right]$ for every day of the year. The latter function depends on how many times a certain day of the year were in our database by specific day of the week. The function is close to periodic with a period of 7 days, while increasing the sample size will tend to a constant.

The public holidays (which are celebrated on the same day every year) significantly affect the statistics (see Fig. 2, 3).

We determine the mean square spline $M_{Year}^{spline}\left[N_{Week}Q_X(t)\right]$ (cubic periodic smoothing spline with a defect=1 and a period *T=1 year*, which provides the smallest mean-square deviation) with the weight $M_{Year}\left[F_{Week}Q_X(t)\right]$ by the values $M_{Year}\left[N_{Week}Q_X(t)\right]$ and with the following knots: 1, 7, 8, 9 January; 19, 22, 23, 26 February; 6, 8, 9, 11, 12 March; 14, 28, 29 April; 7, 8, 11, 15 May; 9, 10, 18

June; 23 July, 4, 18, August; 9, 10 September; 25 November; 7, 26, 30 December. We used periodic boundary conditions for the splines. The graphs of the splines are represented in Figs. 2, 3.

This set of knots optimizes the description of the fluctuations in NAT during the days near the public holidays. Our numerical experiments confirmed that for this choice of knots, the error of our final prediction of NAT is smaller. For the definition of the algorithm and the basic properties of mean square spline, see e.g. Bellman (1974).

The typical NAT $\Psi_X(t)$ for this day of the week and time of the year is calculated by the formula

$$\Psi_X(t) = F_{Week} Q_X(t) \bullet M_{Year}^{spline}\left[N_{Week} Q_X(t) \right].$$

Let us designate, as before, X as a super-group of diseases. To take into account the information about air temperature, we search a periodical spline $K_X(t)$ with a period of 1 year and with the set of knots that was specified above, together with such a function $f_X(T_{\max}, T_{\min})$ of the maximal and minimal air temperatures so that the functional of the deviation took the smallest value:

$$\sum_t \left(Q_X(t) - \left(K_X(t) + f_X\left(T_{\max}(t), T_{\min}(t)\right)\right) F_{Week} Q_X(t) \right)^2 \to \min_{K_X,\, f_X}. \quad (4)$$

We search simultaneously the best (in the sense of minimum (4)) spline $K_X(t)$ in this space and the best function $f_X\left(T_{\max}, T_{\min}\right)$ (in order to describe the contribution of the air temperature into the forecasting of NAT) in the following form:

$$f_X\left(T_{\max}, T_{\min}\right) = c_2\left(T_{\max} + T_{\min}\right)^2 + c_1\left(T_{\max} + T_{\min}\right) + c_0. \quad (5)$$

A further complication of expression (5) for the function $f_X\left(T_{\max}, T_{\min}\right)$ can lead to a decrease of the functional (4), but, as was shown by our additional numerical experiments, this does not reduce the total error of the forecast of NAT (3).

Since the functional (4) depends on sum of $K_X(t)$ and $f_X\left(T_{\max}, T_{\min}\right)$ only, and the constant function is included both in the 33-dimensional space of the periodical splines with respect to t (33 knots) and in the 3-dimensional space of quadratic polynomials on $T_{\max} + T_{\min}$, then we need to

determine 33+3-1=35 parameters. Minimization with respect to these parameters in (4) is produced by the method of least squares.

The typical NAT $\Psi_{X,T_{\min},T_{\max}}(t)$ for this day of the week, time of the year, and temperature is calculated by the formula

$$\Psi_{X,T_{\min},T_{\max}}(t) = F_{Week} Q_X(t) \bullet \left(K_X(t) + f_X\left(T_{\max}(t), T_{\min}(t)\right)\right).$$

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.