ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет «Санкт-Петербургская школа экономики и менеджмента»
Департамент ______________

Мишалкин Иван

Модель прогнозирования потери людей на основе погодных данных/Trees, Mushrooms & Weather - Predictive Model of People's Lost

КУРСОВАЯ РАБОТА

по направлению подготовки 38.03.01 «Экономика»
образовательная программа «Экономика»

Работа сдана в ОСУП
«__» ______________20__г.

Защита состоялась
«__» ______________20__г.

Состав комиссии и подписи:
1.______________________________
2.______________________________
3.______________________________

Оценка за работу_______ (__)
оценка по пяти и (десяти) бальной системе

Студента (-ки) группы №_____
____
____
____
(Фамилия И.О., подпись)

Научный руководитель
____
____
____
(должность, степень, Фамилия И.О., подпись)

Санкт-Петербург
20__г.

Trees, Mushrooms & Weather - Predictive Model of People's Lost

1. Abstract

The aim of this research is in creating people's lost prediction model. It is based on weather data collected over Liningradskaya region, Russia. The number of observations is about 3000 from 2010 till 2015. The paper describes two models: the first one is aimed to prove that the weather has certain impact.And the second has to provide the information, how many people are expected to get lost under certain weather conditions. Relations between weather conditions and peplele's lost has been established, however performance of the second model is rather poor, so extra predictors should be taken into account.

2. Introduction

There are two major organizations in Leningradskaya region, Russia, which deal with search and resque(SAR) operations: emercom(The Ministry of Emergency Situations of the Russian Federation) and Voluntary Rescuers Association "Extremum".

The problem is in shortage of human resources and also in shortage of time: the speed of rescue team reaction greatly impact the chances of survival(Doherty et al. 2014).The time during which the search remains relevant is 51 hours and after this time the chances of survival decline significantly(Adams et al. 2016).

People in Nordic countries often go to forests to collect berries and mushrooms, it is a strong cultural heritage(Salo et al. 2014). Some people, especially retired ones, may get enough income from selling mushrooms and berries(Cai et al. 2011). That's why we assume, that gathering is the main reason for people to go to the forest. The weather influence on the mushrooms emergence and in some cases on peoples' state(Boore & Bock 2016, Jamison et al. 1995)

In order to optimize the usage of human resources the model predicting the probability of people's lost and the amount of such people according to the various weather parameters was made.

3. Emergency search background

The basic demographics of missing people are represented in (Sadeghi et al. 2015) paper. According to this research eldery people are likely to have medical comorbities. The majority who have prior health problems or who are medical dependent mainly consists of people older than 65 years. The age group 45-65 years is under the highest risk of getting injuries. The medical care was needed for 46% of victims, despite the fact that about 90% of victims were in good physical condition.

In Bett's paper it is written about the change in park visits in dynamics(Bett 1954). Since 90's the share of women participating in outdoor activities in the United States of America has increased greatly. the total amount of people, participating in outdoor activity increased by 4.4%, the sum of days of participation in walking for pleasure outdoors grew almost by 14%.

From the SAR data for two year period in (Sadeghi et al. 2015) it was found out that the average age of missing people is 36 years, the majority of incidents involves hiking, the mean temperature is 13°C in addition the wind is present what increases the chances of hypothermia. The temperature varied from -18°C to 35°C.

The relation between SAR operations and weather conditions was set by (Boore & Bock 2016). The authors had data over ten-year period of people who had some health problems and asked for help in Yosemite National

Park. It was found out that the majority of people get lost on clear day(79% of cases) and only 15% when the weather is hot. SAR operations in the Yosmite National Park do not have great differences in the demographics of the missing people from other national parks and wilderness areas.

Some papers set the dependencies between the weather and people's state. (Smedslund & Hagen 2011) found out that the majority of people with rheumatoid arthritis are not weather sensitive, however it is well known that more than 60% of patients with RA believe that their pain is affected by the weather(Jamison et al. 1995) so we can't ignore the rain, pressure and humidity data. The researchers from Tufts-New England medical center found out the relation between barometric pressure, ambient temperature and osteoarthritis pain(Jamison et al. 1995): the pain is higher if the barometric pressure is higher or the temperature is lower. The research of Smedslund and Hagen claims that any weather variable has auto-correlated structure (Smedslund & Hagen 2011). Therefore the weather data should be collected not for the one day.

How often people travel actively and how their behaviour is connected to the weather parameters: Cools and Creemers investigated the behaviour of people that had to choose, how to get to their destination(Cools & Creemers 2012). They used socio-demographic data, transport- and travel-related attributes and also weather data. In some cases people prefer transport to going on foot what in our case may mean that in certain conditions going to the forest may seem unattractive to some people. In their research Cools and Creemers used fog, rain, snow, temperature as a weather data.

4. Data description

The data was collected from "EXTREMUM" database[1]. There is data about SAR operations, initiated in Leningradskaya region(square of 84500 km^2) over a five-year period(since 01/06/2010 till 26/12/2015). We have the date when the person was registered as being lost and the place where the person

[1] http://www.extremum.spb.ru

Table 1. Descriptive statistics on categorical data for logistic regression

	lost_or_not	fog_yesterday	rain_yesterday	thunder_yesterday	three_days_rain
1	0:1,383	0:2,159	0:1,424	0:2,656	0: 883
2	1:1,409	1: 633	1:1,368	1: 136	1: 1,909

tended to go, this called the entrance point.

The other part of data is the weather in the place where the person get lost for the day when the incident happened. We also got the weather for one week before the incident. All weather data was collected from the “Weather Underground“ web-site[2] using API methods. This service gets the information from more than 180000 personal meteostations throughout the world. Moreover the data flows from the automatic meteostations located in the airports. The daily summary was collected and it contains the information about fog, rain, snow, thunder, mean pressure, temperature, dewpoint, wind speed, wind direction, the highest and the lowest temperature of the day and the amount of percipitations.

5. Probability of getting lost

The data was enriched in the assumption that if there is no registration of missing people on certain date this means there are no lost people that day. Having people losts on a certain day we assume that this very day but another year noone missed. After all preparations the available data consisted of 2892 observations. Two tables(table 1, table 2) represent basic descriptive statistics for the data.”Fog_Yesterday” and ”Rain_Yesterday” are categorical variables, which show weather fog or rain were present, ”MeanTempC_Yesterday” is the average temperature of the day, ”MeanPressureMBar_Yesterday” is the average pressure during the day, ”MeanWindSpd_Yesterday” is the average wind speed during the day, ”MeanVisibility_Yesterday” is the average visibility during the day, ”Humidity_Yesterday” is the average humidity during the day, ”Precipitation_Yesterday” is the amount of precipitation during the day, ”Three_Days_Rain” is the categorical variable: if it rained three days through

[2] http://api.wunderground.com

Table 2. Descriptive statistics on continious data for logistic regression

Statistic	Mean	St. Dev.	Min	Max
meantempc_yesterday	12.685	5.530	−24	29
meanpressurembar_yesterday	1,014.393	9.081	978	1,044
meanwindspd_yesterday	8.433	4.182	0	39
meanvisibility_yesterday	9.775	4.448	0	30
humidity_yesterday	77.227	10.069	30	100
precipitation_yesterday	1.322	4.217	0	88
pressure_change	−11.366	68.478	−925.602	42.195
temp_change	0.404	3.054	−12.692	14.769
sum_prec	13.633	34.559	0.000	906.700
maxtempc_yesterday	16.636	6.241	−18	37

two days before the person get lost, "Pressure_Change" and "Temp_Change" show the deviation between the average pressure and temperature during the week and ones during the day of SAR operation, "Sum_Prec" is the amount of precipitation during the week, "MaxTempC_Yesterday" is the highest temperature during the day.

The first model is based on logit regression. This model provides the information about the probability whether people get lost under certain weather conditions. The second part is based on linear model and predicts how many people are likely to become lost under certain weather conditions. The important assumption related to the data is the fact that SAR operation is usually registered on the next day, so the predictors are chosen with this lag taken into account.

For the first model predictors are fog, rain, mean temperature, mean pressure, mean wind speed, mean visibility, mean humidity, mean precipitation, maximum temperature, the fact whether it had rained for three days or not before the day when someone missed, the difference between average pressure during the week and the pressure at the day, when the person missed, the same difference in temperature and the total amount of precipitation during the week before the day, when person missed. All of the predictors turned out to be meaningful: 10 of them at the significant level of promille percent and the precipitation is at the 1 percent significant level(table 6).

The data was divided into two parts: on 80% we trained the model and then tested it on the least 20%. ROC(receiver operating characteristic) curve was also used to evaluate the usefulness of the model(figure 1). This curve shows, in our case, the ratio between the share of people who got lost and our model revealed it(this is called true positive rate) and the share of people that did not get lost, but our model predicted them as being so(this is called false positive rate). The graph is plotted at various threshold settings. True positive rate(TPR) is also known as sensitivity of the algorithm and false positive rate(FPR) is also known as specificity of the algorithm. The higher the square under the graph is the better the model predicts. In terms of the curve the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the model.

The second figure is aimed to show which threshold setting to choose(figure 2). On the x-axis there is cut-off parameter and the plot contains three curves. The solid one represents the general accuracy of the model. It reaches peak

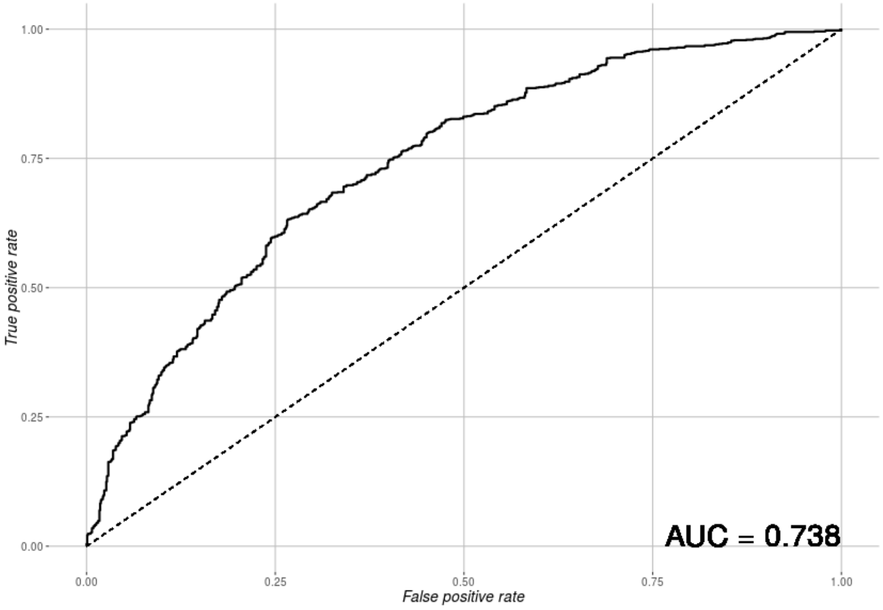

Figure 1. ROC curve
The closer the ROC curve is to the upper left corner, the higher the overall accuracy of the model

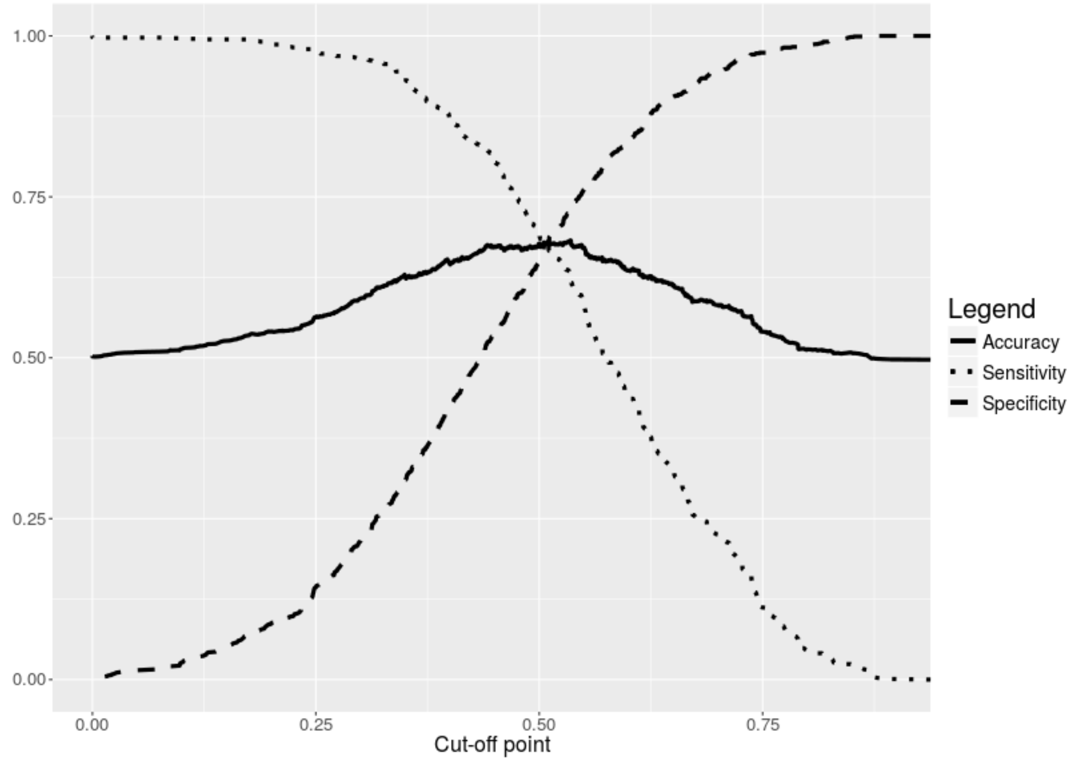

Figure 2. Cut off
Various cut-off points give various ratio of true positive rate and true negative rate

and fluctuate at the same level at the cut-off range of 0.43-0.54. The dotted curve represents sensitivity: the lower cut-off parameter we take the more lost people we predict, however we also will make more mistakes because according to the model there also will be a huge amount of people that didn't lost but the model claims the opposite for them. The dashedd curve specificity. This is also called true negative rate: the share of people who did not lost and the model predicted so. As the accuracy of the model does not change in the cut-off interval of 0.43-0.54 and we would like to get the high-

Table 3. Confusion table for logistic regression

	Not Lost	Lost
Predicted as not lost	122	50
Predicted as lost	151	236

est share of rightly predicted lost poeple so we have to neglect the increasing share of wrongly classified as missed people. Taking all this into account we determine the cut-off level at 0.43. On the test data we get the confusion table(table 3). According to this data the model has the accuracy of 64%.

6. The amount of people that are under risk of getting lost

The second part of this paper is aimed to evaluate how much people will be lost at a particular day in relation to weather data. The data for this model consists of everyday weather observations from the first SAR operation(01/06/2010) till the last one(26/12/2015). After cleaning data 2035 observations remained(table 5). Most of the predictors are similar to ones in the logistic regression model. "presdif" is the difference between the average pressure during the previous week and the pressure the day before SAR operation. "tempdif" is the difference between the average temperature during the week and the day before the SAR operation. "weekend" is a categorical variable which shows whether it is a weekend or a weekday on a particular day. Monday was also considered as a weekday because of the lag between getting lost and SAR operation. "presum" is a total amount of precipitations during the previous week. The data was divided into two parts: the train one(80%) and the test one(20%)

The second part of this paper is aimed to evaluate how much people will be lost at a particular day in relation to weather data. The data for this model consists of everyday weather observations from the first SAR operation(01/06/2010) till the last one(26/12/2015). After cleaning data

Table 4. Confusion table for linear regression

	0	1	10	12	13	18	2	3	4	5	6	7	8	9
0	162	14	0	0	0	0	0	1	0	0	0	0	0	0
1	114	30	1	0	2	1	9	6	4	2	0	3	2	1
2	25	8	1	1	0	0	9	7	2	3	2	0	0	0

2035 observations remained(table 5). Most of the predictors are similar to ones in the logistic regression model. "presdif" is the difference between the average pressure during the previous week and the pressure the day before SAR operation. "tempdif" is the difference between the average temperature during the week and the day before the SAR operation. "weekend" is a categorical variable which shows whether it is a weekend or a weekday on a particular day. Monday was also considered as a weekday because of the lag between getting lost and SAR operation. "presum" is a total amount of precipitations during the previous week. The data was divided into two parts: the train one(80%) and the test one(20%)

Table 5. Data description for the linear regression model

Statistic	N	Mean	St. Dev.	Min	Max
meantempc	2,034	6.705	9.894	−24	29
meandewptc	2,034	2.697	8.806	−27	20
meanpressurembar	2,034	1,013.111	11.058	970.770	1,055.610
meanwindspd	2,034	10.496	4.553	1	33
meanvisibility	2,034	9.046	1.297	0.600	10.800
humidity	2,034	76.881	12.596	31	99
maxtempc	2,034	9.965	10.570	−19	37
mintempc	2,034	3.385	9.476	−29	24
maxhumid	2,034	93.173	6.856	46	100
minhumid	2,034	53.236	20.028	4	93
maxdewptc	2,034	5.441	8.655	−22	24
maxpressurembar	2,034	1,016.563	10.431	982	1,057
minpressurembar	2,034	1,009.687	11.780	967	1,054
maxwindspd	2,034	23.971	12.429	7	223
minwindspd	2,034	2.654	3.174	0	18
maxvisibility	2,034	10.016	0.612	2.400	29.000
minvisibility	2,034	5.218	3.610	0.000	10.000
precipitation	2,034	1.891	20.199	0.000	900.000
mindewptc	2,034	−0.324	9.165	−33	18
presdif	2,034	0.488	12.733	−40.107	155.644
tempdif	2,034	−0.017	3.400	−15.286	12.000
presum	2,034	13.281	53.451	0.000	912.700

The model is based on linear regression. It also includes categorical variables: rain, fog, snow, hail etc . The parameters of the model are given in the table(table 7). The amount of lost people is influenced by fog, thunder, mean pressure of the day, mean visibility, lowest temperature, lowest dew point, the day(weekend or not), difference in pressure and in temperature. To evaluate the model the confusion table on test data will be usedtable 4. Due to this table the model correctly recognized 162 days without any lost people, 30 days when only 1 person has missed and 9 days, when 2 people got lost. The majority of mistakes were made when the model predicted one person missed: 114 false predictions. All in all, there are 301 correct prediction of 410 observations, so the accuracy of the model is about 73%.

7. Conclusion

The main result of this research is the fact that the weather has rather tangible influence on people's losts. The presence of fog, rain, thre days of raining in a raw, increase in preassure, mean temperature, mean preassure, mean visibility, amount of precipitation - all these factors increase the probability of getting lost. Increase in Humidity, mean wind speed, temperature change, summary precipitation and increase in maximum temperature decrease the probability of getting lost. The accuracy of the logit model, built on these predictors is 64%. The other model revealed very few statistically significant predictors, they are: presence of fog, thunder, the increase in pressure, minimum temperature, mean temperature. The linear model is rather poor as it may predict only days without SAR operations.

References

Adams, A. L., Schmidt, T. A., Newgard, C. D., Federiuk, C. S., Christie, M., Scorvo, S. & DeFreest, M. (2016), 'Search Is a Time-Critical Event: When Search and Rescue Missions May Become Futile', *Wilderness & Environmental Medicine* **18**(2), 95–101.
URL: *http://dx.doi.org/10.1580/06-WEME-OR-035R1.1*

Bett, W. R. (1954), 'On trends in medical literature.', *Missouri medicine* **51**(12), 1016–1017.

Boore, S. M. & Bock, D. (2016), 'Ten Years of Search and Rescue in Yosemite National Park: Examining the Past for Future Prevention', *Wilderness & Environmental Medicine* **24**(1), 2–7.
URL: *http://dx.doi.org/10.1016/j.wem.2012.09.001*

Cai, M., Pettenella, D. & Vidale, E. (2011), 'Income generation from wild mushrooms in marginal rural areas', *Forest Policy and Economics* **13**(3), 221–226.
URL: *http://www.sciencedirect.com/science/article/pii/S1389934110001474*

Cools, M. & Creemers, L. (2012), 'The dual role of weather forecasts on changes in activity-travel behavior', *Journal of Transport Geography* **28**, 167–175.
URL: *http://www.sciencedirect.com/science/article/pii/S0966692312002682*

Doherty, P. J., Guo, Q., Doke, J. & Ferguson, D. (2014), 'An analysis of probability of area techniques for missing persons in yosemite national park', *Applied Geography* **47**, 99 – 110.
URL: *http://www.sciencedirect.com/science/article/pii/S0143622813002506*

Jamison, R. N., Anderson, K. O. & Slater, M. A. (1995), 'Weather changes and pain: perceived influence of local climate on pain complaint in chronic pain patients', *Pain* **61**(2), 309 – 315.
URL: *http://www.sciencedirect.com/science/article/pii/030439599400215Z*

Sadeghi, R., Konwinski, J. C. & Cydulka, R. K. (2015), 'Adirondack Park incidents: a retrospective review of search and rescue reports from 2008 and 2009.', *Wilderness & environmental medicine* **26**(2), 159–63.
URL: *http://www.sciencedirect.com/science/article/pii/S1080603214002634*

Salo, M., Sirén, A. & Kalliola, R., eds (2014), *Front Matter*, Academic Press, San Diego.
URL: *http://www.sciencedirect.com/science/article/pii/B9780123972040010019*

Smedslund, G. & Hagen, K. B. (2011), 'Does rain really cause pain? A systematic review of the associations between weather factors and severity of pain in people with rheumatoid arthritis', *European Journal of Pain* **15**(1), 5–10.
URL: *http://dx.doi.org/10.1016/j.ejpain.2010.05.003*

8. Appendix

Table 6. Logistic regression: results

	Dependent variable:
	lost_or_not
fog_yesterday1	2.705***
	(0.109)
rain_yesterday1	1.304***
	(0.096)
meantempc_yesterday	1.146***
	(0.027)
meanpressurembar_yesterday	1.004***
	(0.001)
meanwindspd_yesterday	0.948***
	(0.011)
meanvisibility_yesterday	1.071***
	(0.010)
humidity_yesterday	0.971***
	(0.005)
precipitation_yesterday	1.036***
	(0.011)
three_days_rain1	1.501***
	(0.091)
pressure_change	1.003***
	(0.0005)
temp_change	0.907***
	(0.015)
sum_prec	0.971***
	(0.003)
maxtempc_yesterday	0.853***
	(0.026)
Constant	0.350
	(0.259)
Observations	2,980
Log Likelihood	−1,881.513
Akaike Inf. Crit.	3,791.027
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7. The linear regression: Results

	Dependent variable:
	Losts
fog1	0.576***
	(0.142)
rain1	−0.155
	(0.113)
snow1	−0.039
	(0.143)
thunder1	−0.614***
	(0.220)
meantempc	0.072
	(0.057)
meandewptc	0.033
	(0.071)
meanpressurembar	0.020***
	(0.005)
meanvisibility	0.078*
	(0.045)
humidity	0.011
	(0.013)
mintempc	−0.135***
	(0.037)
maxdewptc	0.033
	(0.040)
precipitation	−0.001
	(0.002)
mindewptc	0.064**
	(0.033)
weekend1	0.255***
	(0.090)
presdif	−0.010**
	(0.004)
tempdif	−0.045***
	(0.014)
presum	−0.001
	(0.001)
Constant	−21.201***
	(5.616)
Observations	1,627
R^2	0.137
Adjusted R^2	0.128
Residual Std. Error	1.784(df = 1609)
F Statistic	15.065***(df = 17; 1609)
Note:	*p<0.1; **p<0.05; ***p<0.01