Demographic Forecasts Based on Queries to Yandex Search Machine

Anna Boldyreva

Moscow Institute of Physics and Technology, Russian Academy of National Economy, Moscow, Russia

anna.boldyreva@phystech.edu

Abstract. *We propose models for indirect forecast of two demographic indicators (registered newborns and marriages) based on user queries to the Yandex search machine. Models are built using the GMDH technique. The resulting accuracy is equal 90-99%. It allows us to recommend the proposed technique as an additional source of information for corresponding governmental institutions.*

Keywords

GMDH, Internet queries, demography

1 Problem setting

Traditionally, demographic predictions use regression models with various variables reflecting economic status and traditions of a given country or region [1,2]. In particularly, it concerns the number of newborns and the number of marriages. We propose to use the other models based on search queries to Internet as independent variables. In our case it could be queries of parents who expect a baby and it could be queries of people who prepare for getting married

2 Data

Dynamics of indicators (newborns and marriages) were downloaded from the portal of the Federal State Statistics Service[1] The data cover the period March 2013 - January 2015. To build the independent variables we used 7 word bases were collected: economic term base – 6468 elements; juridical term base - 4462; base of the articles related to economic crimes of a criminal code - 365; brands and goods base - 3013; bases of positive and negative emotions – 312 and 274; base of slang - 2465; base of lemmas – 18638 2) Then the base of 5 million queries to the Yandex search machine for the year was downloaded and the most frequent 2-8-letter combinations (n-grams, n=2,3,…8) were extracted. Taken together queries and n-grams are named descriptors.

The statistical services of the search engine of Yandex[2] do not give access to the statistics for more than two years, so we collected monthly dynamics of descriptors in Russia from March 2013 to February 2015. Table 1 shows the examples of descriptors. The descriptors having the highest coefficient of correlation (absolute values) with the indicators were used as independent variables in modeling. For the experiments we selected 8 descriptors.

3 Modeling

To build the models we used the tool GMDH Shell containing several GMDH based algorithms [3,4]. In the experiments we studied 4 options: with and without data transformation of descriptors, combinatorial and neural algorithms. Data transformation means here the operation of square root. The quality of results were evaluated by MAPE. Here MAPE stands for the mean absolute percentage error.. Figure 1 shows the process of modeling for one of experiments

[1] Federal State Statistics Service http:// cbsd.gks.ru/

[2] The service of the statistics for queries to Yandex. http:// wordstat.yandex.ru

Tab.1. Examples of descriptors with high coeffitient of correlation.

Indicators	Descriptors with positive coefficient	Descriptors with negative coefficitent
newborns	create, hero, oatmeal, Gorky Park, intern, water, metis, sweating, mugs, got knocked up, cucumbers, carriage	banditry, corps de ballet, teaching, shophelp, printer, drawing, firearms, shipbuilding, Gzhel, nekafe, expensive, sushihit
marriages	hen-party, beach, veil, park, camp, marquee, yacht club, Gagra, anapka, roaming, deck chair, headwaiter	social movement, cossacks, ethanol, hobby, passivity, incoherence, blankets, private enterprise, craft, Internet, needlework, yarn

Tab.2. Values of MAPE for different experiments.

	Neuronet, no roots	Neuronet, roots	Combi, no roots	Combi, roots	AR
Newborns	1,2%	1,8%	1,9%	2,3%	3,5%
Marriages	7,5%	6,9%	9,9%	8,3%	7,7%

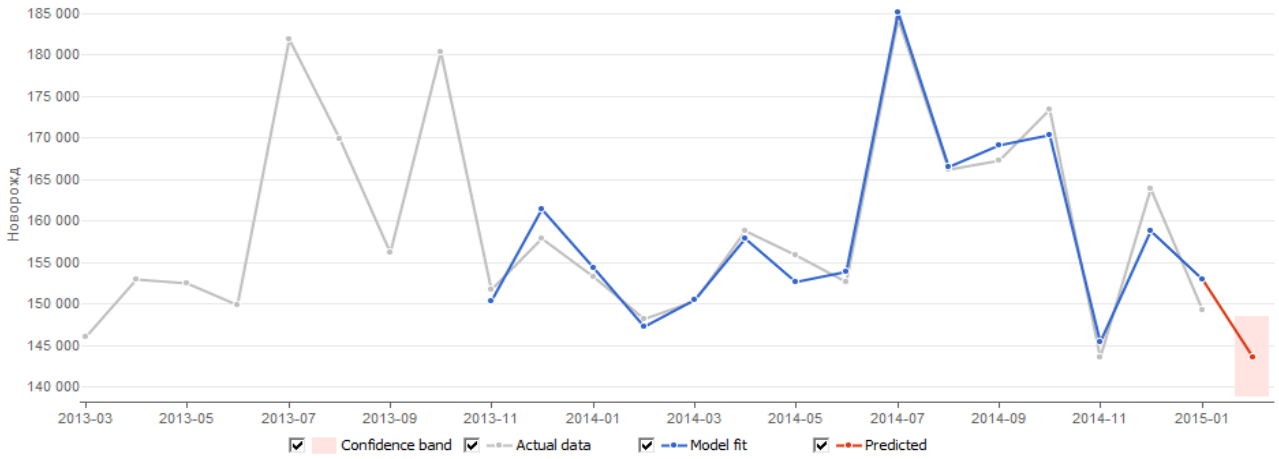

Fig.1. The best model for newborns: neural algorithm without transformation

4 Conclusions

In the paper we build the models for forecasting dynamics of demographic indicators based on the dynamics of queries to Yandex search machine. We study several options of modeling to select the best one. The accuracy of models proved to be higher than in traditional autoregressive models. So, the proposed technology could be recommended for governmental institutions related to demographic studies. In the future we intend to build particular models for forecasting these indicators in Russian regions.

References

[1] Lee R.D.: Forecasting births in post-transition populations: stochastic renewal with serially correlated fertility. *Journal of the American Statistical Association*, 1974, vol. 69, No. 347, pp. 607- 617

[2] Pollard J.: A discrete-time two-sex age-specific stochastic population program incorporating marriage. *Demography*, 1969, vol. 6, issue 2, p. 185-221

[3] Stepashko V.: Ideas of academician O. Ivakhnenko in Inductive Modeling field from historical perspective, *Proc. of 4th Intern. Conf. on Induct. Modeling (ICIM-2013)*, Kyev: NAS of Ukraine, Prague: Prague Tech. Univ., 2013, pp. 31-37

[4] Stepashko V.: Method of critical variances as an analytical tool of the inductive modeling theory. *Journ. of Inform. and Automat. Sciences, Begell House Inc*, 2008, vol. 40, No. 3, pp. 47-58