DISCUSSION PAPER SERIES

IZA DP No. 9503

Predicting Road Conditions with Internet Search

Nikos Askitas

November 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Predicting Road Conditions with Internet Search

Nikos Askitas
IZA

Discussion Paper No. 9503
November 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 9503
November 2015

ABSTRACT

Predicting Road Conditions with Internet Search

Traffic jams are an important problem both on an individual and on a societal level and much research has been done on trying to explain their emergence. The mainstream approach to road traffic monitoring is based on crowdsourcing roaming GPS devices such as cars or cell phones. These systems are expectedly able to deliver good results in reflecting the immediate present. To my knowledge there is as yet no system which offers advance notice on road conditions. Google Search intensity for the German word stau (i.e. traffic jam) peaks 2 hours ahead of the number of traffic jam reports as reported by the ADAC, a well known German automobile club and the largest of its kind in Europe. This is true both in the morning (7 am to 9 am) and in the evening (5 pm to 7 pm). I propose such searches as a way of forecasting road conditions. The main result of this paper is that after controlling for time of day and day of week effects we can still explain a significant portion of the variation of the number of traffic jam reports with Google Trends and we can thus explain well over 80% of the variation of road conditions using Google search activity. A one percent increase in Google stau searches implies a .4 percent increase of traffic jams. Our paper is a proof of concept that aggregate, timely delivered behavioural data can help fine tune modern societies.

JEL Classification: R41

Keywords: stau, traffic jams, highways, road conditions, Google Trends, prediction, forecasting, complexity, endogeneity, behaviour, big data, data science, computational social science, complex systems

Corresponding author:

Nikos Askitas
Institute for the Study of Labor (IZA)
Schaumburg-Lippe-Strasse 5-9
53113 Bonn
Germany
E-mail: askitas@iza.org

1. INTRODUCTION

According to the German automobile club ADAC, in 2014 there have been 475,000 traffic jams on German highways which amounted to 960,000 kilometres of jammed traffic. These numbers represent an increase of 14.4% and 15.6% respectively compared to the year before. This is just an instance of a many year trend which is only expected to get worse even though the report of the ADAC estimates that most of the current increases are due to progress in the method of documentation of traffic jams. According to the ADAC report[1] these traffic jams amounted to 285,000 lost hours which is 32 years.

We can reasonably imagine (see for example Duranton and Turner (2011)) that adverse road conditions of this kind contribute to a host of undesired side effects such as increased pollution, energy waste, additional transportation and production costs, waste of labor, delays in product deliveries and that they also contribute to worsening health conditions as well as to more accidents and even road rage. If we thought of a city, a region, a country or any other social unit as a large living organism road traffic would be one of its circadian rhythms and traffic jams would be an obstruction to its entrainment. It is hence not surprising that obstructing the smooth flow of traffic sends ripple effects deep into many aspects of socioeconomic life. Understanding and forecasting road conditions is important for the benefit of drivers but also obviously for economic reasons.

The emergence of traffic jams is a complex matter and may have many causes. Some are simple and have to do with fluctuating degradation of the available infrastructure (e.g. accidents, road constructions and the like) and some are more complex and depend on behavioural, topological and dynamical complexities. Two well known such complex phenomena are discussed in Braess et al. (2005), where it is shown that expanding the road network may paradoxically worsen road performance and in Sugiyama et al. (2008) where the spontaneous emergence of the so called "phantom jams" are experimentally investigated or in Flynn et al. (2008) where a theoretical model is discussed. A host of other non-linearities are easy to imagine and in fact to reflect on from one's own experience in traffic. For example when a roundabout with four traffic arteries attached to it fails we have a failure of all arteries which cascades radially possibly reaching even more roundabouts. It is due to these complexities that forecasting traffic jams is not a matter of simple linear regression modelling. In fact there is an extensive literature both empirical and theoretical which develops "fundamental diagrams of traffic" i.e. studies the relationships between traffic density, velocity and flow (see e.g Helbing (2009) and the literature therein). Nonetheless we develop such a model which captures a large portion of the aggregate variation of traffic on German roads and we are able to do this 2 hours in advance.

The core observation which sparked this paper is that every morning `stau` searches in Germany peak at 6 am and then they start to dissipate as drivers are being injected into the traffic. Two hours later we have the morning peak of ADAC traffic reports. Similarly at 16:00 hrs every afternoon we observe the search peak and two hours later we see a peak of the ADAC traffic reports. Clearly performing a Google search and driving at the same time are mutually

[1] https://www.adac.de/infotestrat/adac-im-einsatz/motorwelt/Staubilanz2014.aspx

exclusive and this is what creates this advance. The proliferation of mobile applications which allow one to speak searches into a cell phone as well as to graphically get current crowdsourced traffic conditions may diminish the effectiveness of this method but my suggestion is that in addition to reading about current traffic conditions drivers should be taking into account Google search intensity conditions two hours ago. In other words the Google searchers of two hours ago are the drivers in the traffic of now.

Every morning drivers wake up with an origin and a destination on hand. Some for example go from home to work but also from work to home for those doing a night shift. The process of how many people do so every day has some regularities, like 24-hour or 7-day periodicities, but it also has a certain stochastic nature. In order to get from their origin to their destination the drivers need to inject themselves into the road network. When they look for traffic jam information before driving they want to know something about the probability of congestion. This information may be used by some of them who have some leeway to vary the departure time so that they avoid congestion. The phenomenon at hand is reminiscent of a fluid queue in queueing theory. In that sense what drivers are doing when they are searching for traffic jam information is a sort of leaky bucket algorithm (see Kulkarni (1997)). The leaky bucket algorithm is a mechanism which is used to prevent network congestion. Before a data packet enters the network it checks in a token buffer for the existence of tokens. When it finds a token it deletes it and enters the network. When it does not find one it does not enter the network. A regulating mechanism may be dynamically filling the token buffer. When the token buffer is full the token to be added is simply discarded (leaking). What the drivers are doing is to try to see whether there is congestion already and to get a sense of the current state and compare it to their priors. In other words they use current congestion information as their self-administered token buffer. The intensity by which they search in Google would give them a sense of the queue and can hence help them or a central planner build a more efficient token buffer one which reduces congestion.

In this paper I propose a formula for estimating the country wide, aggregate number of hourly ADAC traffic jam reports which has time of the day, day of the week and `stau` search intensity as its explanatory variables. There are currently several systems which provide live road conditions. Such systems use crowdsourcing of volunteers and leverage the fact that such volunteers are roaming members of a digital network and are equipped with Global Positioning Systems able to compute position, direction and velocity of motion[2]. To my knowledge there is no system available which can have as much as two hours advance on the emergence of such road conditions. Our target variable is a crude proxy for road conditions but this paper is a proof of concept that with better target data Google Search can be a powerful predictive tool of traffic conditions. In fact this paper is among very few in the literature of Google Search data which uses hourly data, has a clear behavioural foundation and detects two phenomena with a causal phase difference. Moreover it suggests that the Google Traffic team ought to work closer with the Google Trends team to better forecast traffic. Two hours advance notice

[2]Such systems include Google Traffic, a feature of Google Maps (https://en.wikipedia.org/wiki/Google_Traffic) and the crowdsourcing system of INRIX (https://en.wikipedia.org/wiki/INRIX). More providers can be found here: https://en.wikipedia.org/wiki/Traffic_reporting.

is an enormous amount of time when it comes to traffic jams.

I discuss the ADAC data and the Google Trends Data in Section 2 where I also run some descriptive tests. In 3 I do a forecasting exercise and in the last Section 4 I draw some conclusions from this exercise.

2. DATA

On September 28 I started collecting ADAC traffic jam data programmatically every five minutes from the website[3] of the German automobile club ADAC. Each observation consists of a timestamp, region name and current count. I use the aggregate average hourly count as my target time series. Figure 1 depicts some diagnostics for this data. As expected it has 24 hour and 7 day periodicities, autocorrelation and, due to non-linearities and endogeneities, it is leptokurtic and fat-tailed.

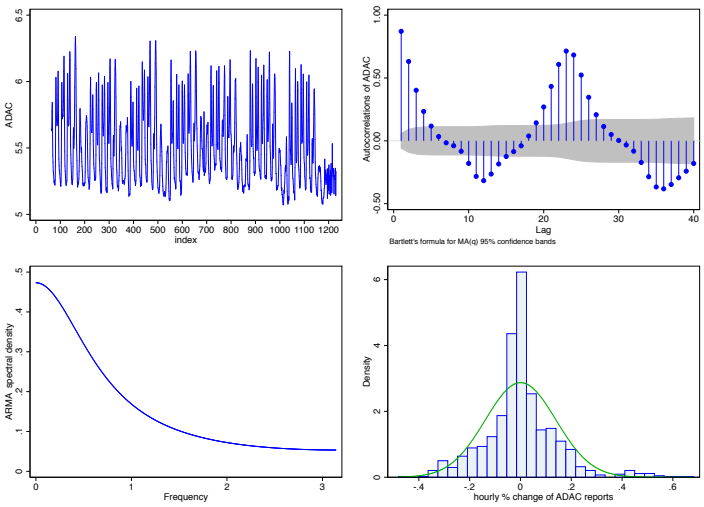

FIGURE 1.— The natural logs of the hourly number of ADAC traffic jam reports (top left), its autocorellations (top right), spectral density (bottom left) and histogram of percentage changes (bottom right). The series exhibits a significant amount of autocorellation, a strong circadian rhythm with day of week and hour of day fixed effects and a unimodal leptokurtic and fat tailed distribution of hourly percentage changes.

Data Source: ADAC (adac.de) and own calculations.

The Google Trends data are hourly data of searches containing the word `stau` which is German for traffic jam. I am discussing some aspects of this data in Askitas (2015). Before before moving on to discussing the hourly data I dissect such searches in order to provide support for the identification strategy. In this way we demonstrate that our searches are made by drivers and hence our method does not just produce spurious results. Figure 2 shows weekly

[3] https://www.adac.de/reise_freizeit/verkehr/aktuelle_verkehrslage/default.aspx

searches which contain the word "stau" since 2004[4] together with a reduced series. The trend in this figure is in agreement with the annual ADAC traffic jam reports.

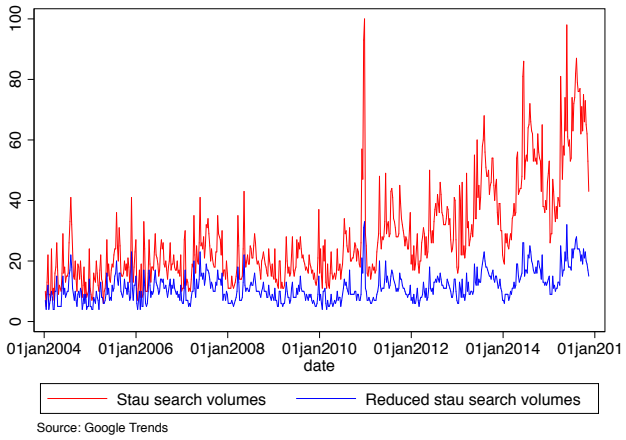

FIGURE 2.— The reduced series are searches containing the word stau without containing any of: `nrw, a7, a2, a3, a1, wdr, a8, a5, aktuell, autobahn, a9, info, swr3, bayern, hamburg, a4, staumelder, adac, berlin, a6, verkehr, ffh, hessen, köln, münchen, swr, a81, a61, deutschland`. A 60% of the total stau search volume is accounted for by these words.

Data Source: Google Trends and own calculations.

The reduced series is constructed by subtracting those searches which contain any of the top 30 additional words. We see that these 30 words account for 60% of the total volume on the average. These keywords are city or region names (NRW, Bayern, Hamburg, Berlin Hessen, Kln, Mnchen Deutschland) or names of highways (A7, A2, A3, A1, A8, A5, A9, A4, A6, A81, A61). Such searches even reveal which highway the driver will be driving on. We also have radio stations (WDR, FFH, SWR) and other websites such as the one of ADAC. Although we cannot subtract them from the volume due to limitations from the Google Trends data provisioning system we know a further set of keywords which are contained in the reduced series. These keywords further support the thesis that these are prospective drivers. They are: `a45, stuttgart, a40, elbtunnel, frankfurt, a44 stau, meldungen, hr3, a14, app, hannover, online, a24, niedersachsen, ndr, bremen, a46, a10, staumeldungen, aktuelle, karlsruhe, wdr2, a31, a57, a43, baden württemberg, nürnberg, bw, nachrichten, antenne, dresden, b10 stau, a7, a96, österreich, sachsen, bonn,a5`. The highway numbers contained as well as the city and region names imply that our method can predict not just the nation wide aggregate road conditions but also regional conditions. The largest portion of stau searches contains are those that also contain the string "NRW" which stand for Nord-Rhein-Westfalen the region with the densest highway network in Germany as can be seen by inspecting Figure 3.

[4]The enormous spike on December 2010 is due to an extraordinary traffic jam in Germany due to snowfall.

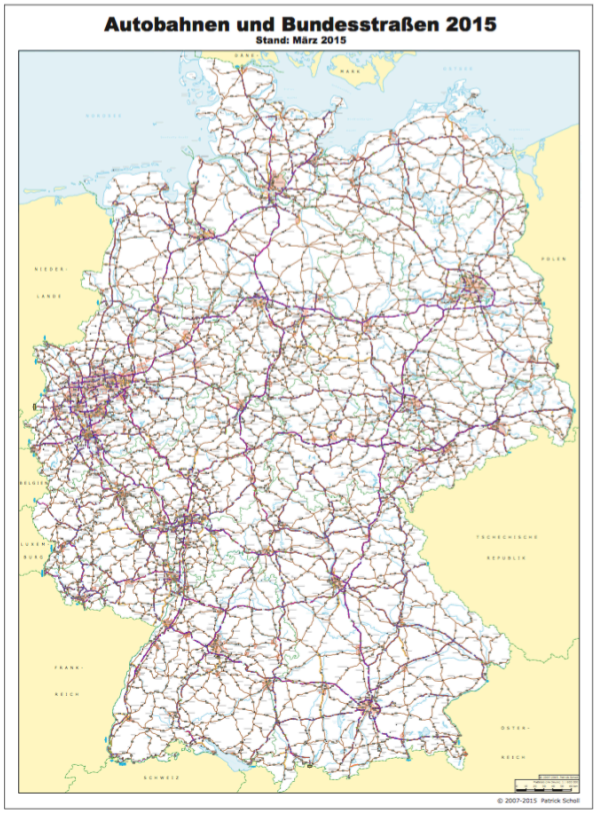

FIGURE 3.— The region with the most stau searches is NRW which also has the densest highway network.

Note: Reprinted here from www.autobahnatlas-online.de with the kind permission of Patrick Scholl.

To recap looking at Google `stau` searches not only parsimoniously identifies soon-to-be members of traffic it also locates the highway they will soon be on. Having established that these searches are performed by future drivers let us now turn to hourly such searches. The characteristics of this series can be seen in figure 4.

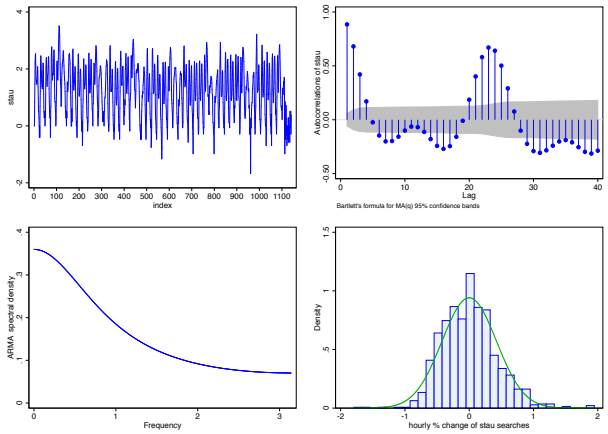

FIGURE 4.— The natural logs of the hourly stau search intensity (top left), its autocorellations (top right), spectral density (bottom left) and histogram of percentage changes (bottom right).

Data Source: Google Trends and own calculations.

The core graph which establishes that Google `stau` searches have a two hour advance on the number of ADAC traffic jam reports is Figure 5

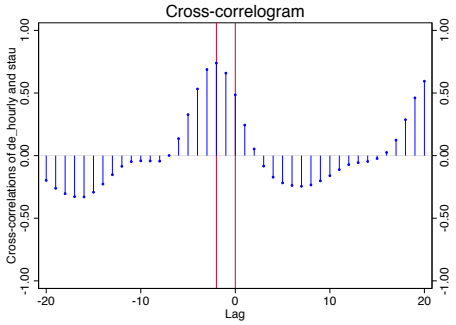

FIGURE 5.— A cross correlogram between the hourly number of ADAC traffic jam reports and the hourly Google search intensity for `stau` establishes that Google search has a two hour advance on road conditions.

Data Source: Google Trends, ADAC and own calculations.

3. FORECASTING

Informed by the observation in Figure 5 I estimate a simple model as follows:

$$S_t = \alpha D + \beta H + \gamma G_{t-2} + \delta + \epsilon_t, \tag{1}$$

where S_t is the natural log of the number of ADAC traffic jam reports at time t (hour), D is the day of the week, H is the hour of the day and G_{t-2} is the natural log of the Google search intensity at time $t-2$. We estimate the parameters $\alpha, \beta, \gamma, \delta$ and ϵ is the error term. I benchmark this model against the autoregressive baseline model:

$$S_t = \overline{\alpha} D + \overline{\beta} H + \overline{\gamma} S_{t-2} + \overline{\delta} S_{t-24} + \overline{\zeta} + \overline{\epsilon}_t, \tag{2}$$

The regressions are restricted to the fixed effects from days and hours that are statistically significant. Table I summarises the results of five OLS regression. The first two models are the simplest possible where we respectively regress the natural log of the number of ADAC reports on the (natural log of) its second lag and the (natural log of) the second lag of Google search intensity for `stau`. Clearly the second model beats the first in terms of RMSE, AIC and R2 and the last model beats all others on all counts, among them the benchmark model (fourth).

One can clearly not do much better than that since I only have a crude aggregate measure of road conditions but this analysis clearly demonstrates that Google Search intensity for `stau` contains advance information on the number of traffic jams two hours before they occur. Better data would most likely allow better models to come to fruition. Figure 6 contains scatter plots and regression lines for the fourth and fifth models.

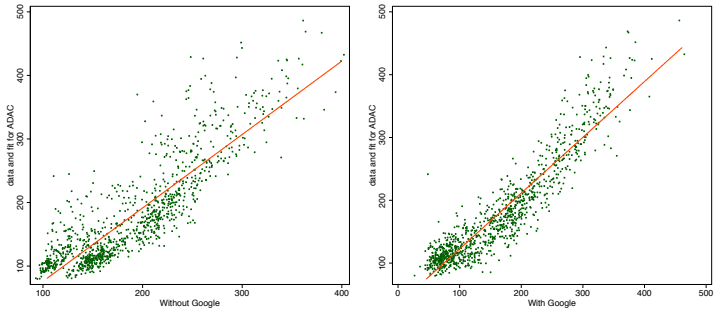

FIGURE 6.— Scatter plots and regression lines for models four and five of table I. Clearly the regression line in the fourth model fails to capture the data optimally. There is obviously a higher degree term.

Data Source: Google Trends, ADAC and own calculations.

It becomes apparent that a second degree term is present in the fourth model. This is not surprising since the probability of pairwise interaction increases with the number of active drivers but reaches a saturation point after a certain critical value is reached. Figure 7 shows a third degree polynomial fit.

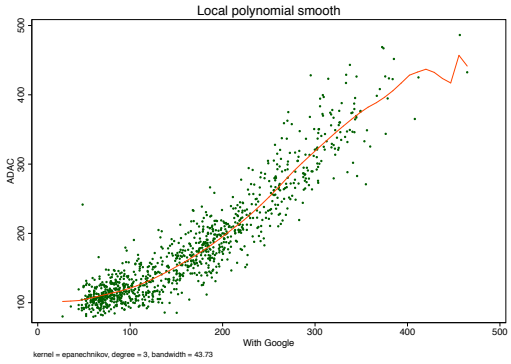

FIGURE 7.— A third degree polynomial fit for data and fit from the best model
Data Source: Google Trends, ADAC and own calculations.

4. CONCLUSIONS

Google search intensity for the word `stau` is an elegant and parsimonious way to capture driving intent in the near future in Germany and hence to predict road conditions two hours in advance. I have demonstrated this using an admittedly crude proxy for road conditions. Better target data is needed and more research is necessary to operationalise the results of this paper. Since Google searches for `stau` are often accompanied by city, region or highway information access to better data such as the data of Google Traffic or other traffic jam information providers (with geographic information attached) could allow one to build predictive systems and in fact even preventive systems. Google Traffic has access to the GPS information of some Android users and not on the universe of all drivers. For the purpose of describing current road conditions a small number of drivers suffices. Google search on the other hand captures driving intent of a much larger portion of the population and does so before road conditions emerge. With sufficient access to Google Traffic and Google Search data it appears as though one should be able to forecast Google Traffic with Google Trends.

Figures 8 show fours major highways (A1, A2, A3 and A7 respectively) and plots the cities from which searches for `stau A1`, `stau A2`, `stau A3` and `stau A7` come from. One can see how these are nicely aligned along the respective highways providing support for the claim that such data can be used for regional traffic forecasting. An interesting accidental observation emerges from this graphic. The town Glauchau appears in the top position for all highways which leads to the conjecture that something beyond and above end user behaviour happens there. Very close to the town is Volkswagenwerk Zwickau the well known automaker. It would appear as though this automaker may already be looking into this type of data.

Regarding the big picture this paper is yet another indication that social science in the upcoming future will indeed look and feel more and more like "doing physics with particles that have feelings"[5].

[5] *"Imagine how much harder physics would be if electrons had feelings!"* – Richard Feynman, Caltech graduation ceremony

REFERENCES

Askitas, N. (2015): "Google search activity data and breaking trends," *IZA World of Labor*.

Braess, D., A. Nagurney, and T. Wakolbinger (2005): "On a paradox of traffic planning," *Transportation science*, 39, 446–450.

Duranton, G. and M. A. Turner (2011): "The Fundamental Law of Road Congestion: Evidence from US Cities," *American Economic Review*, 101, 2616–52.

Flynn, M. R., A. R. Kasimov, J.-C. Nave, R. R. Rosales, and B. Seibold (2008): "On" jamitons," self-sustained nonlinear traffic waves," *arXiv preprint arXiv:0809.2828*.

Helbing, D. (2009): "Derivation of a fundamental diagram for urban traffic flow," *The European Physical Journal B*, 70, 229–241.

Kulkarni, V. G. (1997): "Fluid models for single buffer systems," *Frontiers in queueing: Models and applications in science and engineering*, 321, 338.

Sugiyama, Y., M. Fukui, M. Kikuchi, K. Hasebe, A. Nakayama, K. Nishinari, S. ichi Tadaki, and S. Yukawa (2008): "Traffic jams without bottlenecks?experimental evidence for the physical mechanism of the formation of a jam," *New Journal of Physics*, 10, 033001.

FIGURE 8.— Searches for `stau A1` (marked with a rombus), `stau A2` (marked with a circle), `stau A3` (marked with a square) and `stau A7` (marked with a star) are aligned along the respective highways A1, A2, A3 and A7. This fact supports the thesis that Google search can predict regional traffic conditions especially with access to the actual number of searches instead of to the share of such searches.

Note: Snapshots from Google Maps.

TABLE I

FORECASTING THE NUMBER OF ADAC TRAFFIC REPORTS

	ADAC coef./p-value	ADAC coef./p-value	ADAC coef./p-value	ADAC coef./p-value	ADAC coef./p-value
L2.ADAC	.651***			.391***	
	(.000)			(.000)	
L2.STAU		.366***			.415***
		(.000)			(.000)
D=0			.000	.000	.000
			(.)	(.)	(.)
D=1			.433***	.283***	.268***
			(.000)	(.000)	(.000)
D=2			.465***	.306***	.337***
			(.000)	(.000)	(.000)
D=3			.527***	.349***	.382***
			(.000)	(.000)	(.000)
D=4			.559***	.376***	.399***
			(.000)	(.000)	(.000)
D=5			.480***	.319***	.154***
			(.000)	(.000)	(.000)
D=6			.076*	.032	–.051*
			(.033)	(.338)	(.023)
H=7			.000	.000	.000
			(.)	(.)	(.)
H=8			.437***	.432***	.253***
			(.000)	(.000)	(.000)
H=9			.565***	.476***	.342***
			(.000)	(.000)	(.000)
H=10			.402***	.143**	.304***
			(.000)	(.004)	(.000)
H=11			.294***	–.015	.344***
			(.000)	(.778)	(.000)
H=12			.293***	.047	.424***
			(.000)	(.339)	(.000)
H=13			.323***	.120*	.435***
			(.000)	(.012)	(.000)
H=14			.383***	.180***	.423***
			(.000)	(.000)	(.000)
H=15			.469***	.255***	.466***
			(.000)	(.000)	(.000)
H=16			.562***	.326***	.492***
			(.000)	(.000)	(.000)
H=17			.706***	.433***	.531***
			(.000)	(.000)	(.000)
H=18			.778***	.470***	.552***
			(.000)	(.000)	(.000)
H=19			.651***	.286***	.448***
			(.000)	(.000)	(.000)
const.	1.776***	4.632***	4.509***	2.816***	3.985***
	(.000)	(.000)	(.000)	(.000)	(.000)
Adj. R^2	.423***	.622***	.612***	.672***	.851***
AIC	554.056	85.647	–57.311	–155.917	–639.008
RMSE	.310	.251	.227	.209	.141
No. of cases	1108.000	1108.000	611.000	609.000	609.000

ADAC