New data sources for statistics: Experiences at Statistics Netherlands

Piet Daas[1*], Marko Roos[1], Chris de Blois[1], Rutger Hoekstra[1],
Olav ten Bosch[1], and Yinyi Ma[2]
[1]Statistics Netherlands, [2]Erasmus University Rotterdam, *e-mail: pjh.daas@cbs.nl

Abstract

Apart from the traditional sources used by National Statistical Institutes, like sample surveys and administrative sources, nowadays more and more electronic sources of information are available that potentially can be used for the production of statistics. At Statistics Netherlands we studied four 'new' secondary data sources for this purpose: i) Product prices on the internet, ii) Mobile phone location data, iii) Twitter text messages, and iv) Global Positioning System (GPS) data and traffic loop information. In this paper, for each of these data source an overview is given of the usability of the collected information, as well as the practical and methodological challenges that lay ahead. Some of these sources could be used immediately in the traditional statistical process and for some this is clearly more of a future topic. As for the methodological challenges: these were found to be remarkably similar.

Keywords: New data sources, statistics, data collection

1. Introduction

Traditionally, sample surveys are used by National Statistical Institutes to collect data on persons, businesses, and all kinds of social and economical phenomena. During the last 30 years, more and more statistical institutes have gradually been replacing survey data with administrative data. This shift is mainly caused by the wish to decrease the response burden on the data providers and the desire to produce statistics of sufficient quality in a cost efficient way (Bethlehem, 2010; Snijkers, 2009). Apart from administrative data sources there are, however, also other sources of secondary information available in the world around us that could -potentially- provide data of interest for producers of statistics (Roos et al., 2009). Nowadays, more and more information is processed and stored by many of the ubiquitous electronic equipment surrounding us and the ever increasing use of the internet causes more and more persons (and companies) to leave their digital footprint on the web (Dialogic et al., 2008). All of these sources of information could assist in the production of statistics in the same way as administrative data, but could even provide information describing new social and economical phenomena!

1.1 Goal of this paper

To get a grip on the practical and methodological implications, several 'new' secondary data sources were investigated at Statistics Netherlands. The results of these studies are

described in this paper. Every data source was studied with a potential statistical application in mind. The data sources discussed in this paper are: i) Product prices on the internet, ii) Mobile phone location data, iii) Twitter text messages, and iv) Global Positioning System (GPS) data and traffic loop information. Some of these sources could be used immediately in the traditional statistical process and for some this is clearly more of a future topic. In the remainder of this paper, for each data source, an overview is given of the data collected and the results obtained so far, followed by a discussion on issues identified and challenges remaining. The paper ends with an overall comparison of the findings for the sources studied.

2. Product prices on the internet

The internet, especially the World Wide Web, is a very interesting source of pricing information. At Statistics Netherlands the Consumer Price Index (CPI) department is already gathering data from the web on articles such as airline fares, books, CDs, and DVDs. The data is collected by statisticians that visit websites, look for the relevant data, copy it, and store it into a local database. Special protocols have been developed per CPI-item to ensure that statisticians collect the data using a consistent and structured statistical procedure (Hoekstra et al., 2010). This information could also be automatically collected -at more frequent intervals- by programs that are popularly referred to as web crawlers, web scrapers, or (internet) bots.

2.1 Data collection

There are several ways of automatically collecting data from web pages. The technical ways considered in this study were: with an open source script language (such as Perl and Python), via search engine software augmented with additional scripts, and by using (commercially) dedicated robot tools. Each approach has its pros and cons; see Hoekstra et al. (2010) for more details. Minor disadvantage of the technical ways considered is the inability to collect prices from websites that have information embedded in animations (e.g. 'flash') or pictures. Major disadvantage of search engines is that they are essentially designed to index rather than collect data and therefore need additional scripting to actually gather the data. Used in this way, this approach very much resembles the scripting solution. Search engines were therefore excluded. In house, Perl scripts were written and the commercial software tool Djuggler was used for daily data collection from six web sites for a period of 10 months. Data was automatically collected from four airline websites (already included in the CPI manual data collection process), one housing website, and a website of a chain of unstaffed petrol stations. The data was collected during the night, to reduce the burden on the website, and the scripts and tool used all included a Statistics Netherlands identification string.

2.2 Results

Major problems observed with the scripts and tool during the data collection period were the interaction with dynamic web pages (i.e. web pages generated by a web application), response time of the website (delays in requests), and the occasional change of websites. The first two problems have to be solved during the design phase of the script or tool. It

is the occasional change in the design of a website that affects automated data collection the most. It was found that minor (cosmetic) changes to websites hardly affected the scripts and tool. Major redesigns of a website, however, seriously influenced the data gathering process resulting in wrong or no data at all. In these cases it took considerable time to rewrite the script or reprogram the tool. Major redesigns therefore seriously influence the desirability and costs of using these techniques. During the data collection period, three of the four airline websites changed in such a way that they affected automated data collection; one site even changed twice. Depending on the size of the site change and the level of complexity of the site, rewriting the scripts took between 8 and 40 hours, respectively. The script developed for the airline site that did not change considerably, ran flawless during the whole data collection period.

2.3 Discussion

The results obtained revealed several problems associated with automatic data collection. The first is the fact that automatic collection lacks human checks. To prevent the collection of wrong data, 'robust' scripts need to be developed. We also found that, if automated methods are used to exactly replicate the data collection procedure, it is rarely more cost-efficient. Based on these experiences, we conclude that the main advantages of automatic data collection from websites are not replacing the existing manual collection processes but are i) to download more data (leading to increases in quality, frequency, and speed) and ii) to implement new areas for which manual data collection is very labour-intensive. Examples of the latter are housing prices and fuel prices of a large number of petrol stations.

An intriguing example of the methodological value of downloading large amounts of data is reported in Hoekstra et al. (2010). In figure 1 the results of the daily tracking of airline tickets prices, from a single website, to four European destinations are shown. Collection started 116 days before the day of departure and prices are compared to the average fare

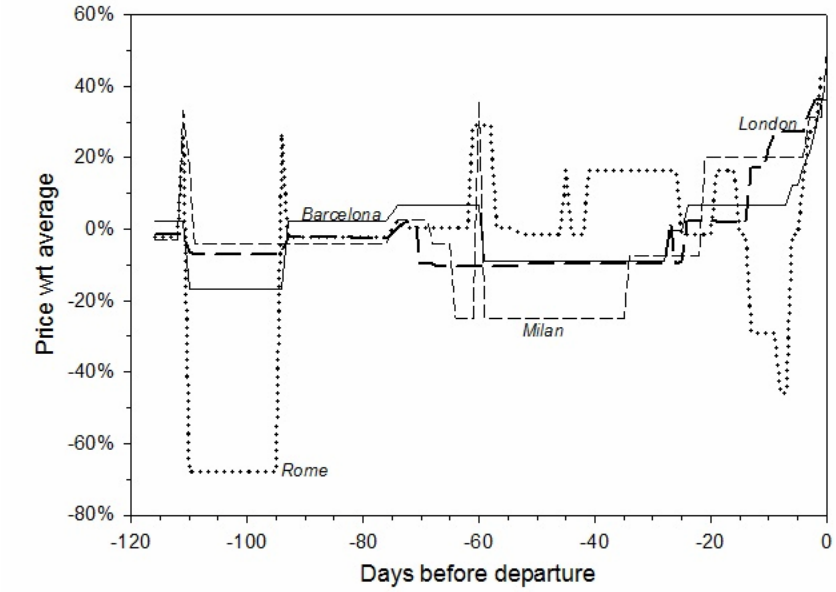

Figure 1. Daily tracked airline ticket prices to four destinations for a 116 day period before departure day. Prices are compared to the average price for each destination.

price for each destination over this period. Figure 1 shows that booking a flight on the last day before departure costs about 40% more than the average price in the data collection period. This premium is fairly consistent for all destinations. The results also reveal that flight prices can fluctuate tremendously, from up to 70% below the average price (Rome, around 100 days before departure) to about 30-40% above average (Rome and Milan, 60 days before departure). The volatility of the data shown in the graph raises the methodological question on how to collect data in order to estimate the "real" price of each flight. The current CPI-method (manually collecting ticket prices three times a month for a date one day before the data of departure) may result in unreliable estimates. Perhaps this type of reliability problem could benefit from the methodological advancements in the scanner data index theory which also involves very volatile and frequent data (Hoekstra et al., 2010). The third issue is the representativity of the data on the website (Ossen et al., 2010). Not all websites display the prices of all products sold in the shop. Also prices may depend on the purchasing channel: the prices for the same products on the website and in the shop might not be identical. Consumer electronics is an example for which this is a serious issue.

3. Mobile phone location data

The use of mobile phones nowadays is ubiquitous. In the Netherlands, around 92 percent of the people use mobile phones on a regular basis. A lot of those people carry their phone with them all day long and use it many times during the day. Every time someone communicates via his/her mobile phone, a signal is send to a nearby located mobile phone mast. Each mast has an unique 'cell' identification code. This cell code provides a source of location information. Much of the activity associated with handling phone traffic is logged by the mobile phone company because it is used for billing information and network optimization. From a statistical perspective, this data could provide information on the geospatial activity of mobile phone users during the day. Examples are the day-time location and movements of people and the social and economic properties of the area covered by phone masts.

3.1 Data collection

For our study a dataset from a telecommunication company in the Netherlands was obtained. The company operates about a third of all mobile phone communication in the Netherlands. The dataset contained records of all call-events, i.e. the sending and receiving of calls and text messages, on their Dutch network for a period of two weeks. Each call-event contained information on the starting time and date, serving cell (the masts 'location'), type of event (sending/receiving of call or text message), and an anonymized unique identification key associated with the phone of the caller. The anonymization process used assured that the identification key for each phone remained the same in the dataset but prevented the identification of individual mobile phones and/or phone numbers. In addition, a cell plan containing information on the geo-locations of the masts used by the telecommunication company was provided. Additionally, we transformed this cell plan into non-overlapping areas covering the Netherlands (Roos et al., 2010).

3.2 Results

The first goal of the mobile phone study was to see if and how the geospatial activity of mobile phone users changed during the day. In this study we assumed that the location and movement of a phone is closely related to the behaviour of the person using it. Another important assumption is that a mobile phone always connects to the nearest available cell within reach and that the areas served by the cells do not overlap. Based on these assumptions, the location and movement of an (unidentified) person can be deduced by determining the serving area of the cell to which the phone connects. For the two week period included in the dataset, the activity (number of calls) in five minute frames was determined in the area served by each cell. Because phones can have more than one call-event in the selected time frame, multiple calls for the same phone were reduced to one. To distinguish the levels of activity in the areas, a logarithmic scaling was used; this was needed because the extreme high phone activity in the 'Randstad' area (the area including the major cities Amsterdam, Rotterdam, The Hague, and Utrecht) would otherwise dominate the activity in the rest of the country. This resulted in a very interesting geospatial animation of which eight frames are shown in figure 2.

The frames (and the animation) reveal that high call intensity areas coincide with areas of high population density; e.g. major cities. More detailed studies revealed that -for most areas- the call activity during the day follows a characteristic repetitive pattern. As a rule, activity is low at night, rises rapidly in the morning (starting around 7 am), displays a lunch dip, an 'I'm coming home' peak at 6 pm, and an 'I'm going to sleep' peak at 10 pm. The activity on working days is usually higher than on weekends and holidays. The activity pattern of some specific areas deviates from this general pattern depending on the location and day of the week. Areas in the centre of 'Randstad' cities have a relative high night time activity and display a much higher day time activity on specific days; e.g. on Sundays when the shops are open and on national holidays such as Queens Day. The pattern also differs for industrial/commercial areas. Especially on weekends and holidays, the call-activity in those areas is much lower. This suggests that commercial/industrial

Figure 2. Call intensity in the Netherlands during a working day.

areas can be identified in this way. The studies did not reveal a commuting effect. More in-depth studies of individual callers might unveil more.

3.3 Discussion

Although the data used in this study already reveals very interesting possibilities for the use of phone data, it also has clear limitations. In the study only the data of a single Dutch telecom provider was analyzed thereby ignoring almost two-third of the mobile phone users in the Netherlands. We also have no knowledge on the chance that the call-event of an individual is included in the dataset, i.e. of the representativity of the data in the set. Additional information on mobile phone users needs to be gathered in order to investigate the representativeness of the mobile phone data for the Dutch population as a whole. Ideally, population density should be deduced from call activity. Current research efforts focus on the use of telecommunication data to estimate the day-time population in specific areas and the day-time movements of people. Other useful purposes could be estimating mobility, tourism, and economic activity.

4. Twitter text messages

It is common knowledge in social sciences that measuring opinions, attitudes, and sentiments is difficult. Recent advances in internet technology may offer an alternative to traditional survey sampling. More and more individuals publish personal information, often including opinions and sentiments, on the internet. Facebook, LinkedIn, and Twitter are examples of social media sites were such data are available. All three media employ the concept of social networks: individuals connect with each other and exchange information. In this study we focus on Twitter. Twitter is a micro blogging service that allows the sending of short text messages of 140 characters in length, called ‘tweets’, either to the general public or specifically to friends, identified as ‘followers’. Public messages are available to all, i.e. also to readers who are not a member of the senders network. Especially the latter feature made Twitter attractive for our study. Because there is a lot of free software available to send tweets from a multitude of devices, such as personal computers, tablets, and mobile phones, an increasing number of people are using this medium. Twitter users are uniquely identified by there user id; usernames can be changed. Apart from a username, a person who creates a Twitter account may also provide a short biography, their location, and a picture. Typical for tweets is the use of ‘hashtags’; words preceded by a ‘hash’ or ‘number’ symbol (#). Hashtags are used to identify the key-words in a message thereby enabling the categorization of messages on Twitter. To study the content and the usability of the information exchanged on Twitter, we collected and analyzed messages. The primary goal of this study was the identification of the topics discussed on Twitter in the Netherlands. From this it was assumed that the amount of messages relevant for official statistics and the area(s) of potential use could be deduced.

4.1 Data collection

After comparing several methods of collecting twitter messages, it was decided that the best approach was to collect messages in an user oriented way. With this approach we

were more certain that topics discussed by only a few individuals would also be obtained. Data collection therefore started by gathering twitter user ids for as many users as possible of which the location information included the words Netherlands, 'Nederland', Holland, the name of a Dutch municipality, or a combination of these. At regular intervals the location information collected was manually checked. Users with location information of obvious non-Dutch locations (such as: 'Holland, Michigan' or 'Amsterdam, Missouri') were manually removed. For every new user id found, all its followers were checked for new user ids. After considerable time and effort, a total of 380,415 unique user ids of apparent inhabitants of the Netherlands were obtained. Next, for each user id, its most recent twitter messages, up to a maximum of 200, were collected and stored in a database. For 61% of the users messages were obtained, the other users either had never written a message or had only written non-public messages to friends. In this way, a total of 12,093,065 public tweets were collected. Apart from its content also additional meta-information was obtained. The latter enabled the identification of a message as authentic, a forwarded version (a so-called 'retweet'), or as a reply to another message. The meta-information also indicated whether the message included a username, a hashtag, and/or a link to a webpage.

4.2 Results

During the initial analysis of the content of the twitter messages, it became clear that a huge diversity of topics was discussed via this medium. The impression also emerged that for a considerable part of the messages it would be very difficult to identify the topics actually discussed. We therefore first focused on messages containing hashtags; here users already indicated the key words in their message. In our dataset 1,750,074 messages (14.5%) contained one hashtag, 12,378 contained two or more (0.1%). Because of their small number and potential disturbing effect, the latter group was ignored. As users are free to use and introduce hashtags, a considerable number of unique hashtags (16,439) occurred. The distribution of the number of messages for each hashtag was found to be highly skewed; the 77 mostly used hashtags comprised a quarter of the total number of hashtag containing messages. Whereas, the 500 frequently used hashtags represented nearly 50% of all messages. By manually grouping the 500 mostly used hashtags into similar topics, a first overview was obtained of the subjects discussed (in hashtag containing messages) on Twitter in the Netherlands (Table 1). A considerable 'other' group was found (38%) containing messages with unrelated and often unclear topics. This group increased when the remainder of the hashtag containing messages was added, without changing the ratio between the 12 topics already identified. When all hashtag containing messages were grouped, the ratio between the 12 topics already identified remained nearly identical to those obtained from the top 500 hashtags (Table 1). In the end, the 'other' group had increased to 72% of all hashtag containing messages. To get an idea of the topics discussed in the 10,330,613 non-hashtag containing messages, text mining techniques were used to classify these messages with the categorized single hashtag containing messages as a training set and the hashtag as a label. The software program LingPipe was used with an implementation of the DynamicLMClassifier. First results seem to confirm our hashtag-based findings. This suggests that around 5% of the Twitter messages collected could be of potential interest for statistics. Especially politics and events are areas were Twitter could potentially be used to obtain information on the

Table 1. Classification of Twitter messages collected according to the hashtags used.

Category	Description	Examples	Top 500# only (%)	Top 500# no Other (%)	All# no Other (%)
Twitter	Twitter/internet specific language & slang	#durftevragen, #fail, #twexit	12	19	19
Sports	Sports, clubs, and sports events	#WK2010, #ajax, #oranje	9	14	14
Applications	Twitter specific programs	#nowplaying, #lastfm, #in	8	13	12
Politics	Political debates, leaders, and parties	#tk2010, #NOSdebat, #formatie	7	11	11
TV	Dutch TV-programs (no political & no news)	#dwdd, #ohohcherso, #tvoh	6	10	11
Emotions	Sentiment and feelings	#moe, #LOL, #zucht, #heerlijk	6	10	10
Locations	References to a location or municipality	#amsterdam, #utrecht	3	5	5
Products	Referring to products	#iPhone, #iPad, #android	3	4	4
Events	Non-sport and non-political happenings	#twibbon, #LL10, #lowlands	3	4	4
News	Referring to news programs	#nos, #pownews, #Nujij	2	4	4
Companies	Referring to companies	#ns, #google, #tmobile, #KPN	2	4	4
Radio	Dutch radio programs	#3fm, #53j8, #radio1	1	2	2
Other	Rest group, mostly unrelated tags	#koffie, #goedemorgen	38	-	-

opinions, attitudes, and sentiments in the Netherlands. Other possible areas of interest are social and cultural participation and social cohesion.

4.3 Discussion

Although our first results reveal what topics are discussed on Twitter in the Netherlands, it is very difficult to relate this to the (opinion of the) Dutch population as a whole. This is caused by two reasons. First, the results obtained were derived from a classification of (single) hashtag containing twitter messages. Although this set almost comprised around 15% of the total number of messages obtained, results could be biased because i) only messages on particular topics contain hashtags and ii) only a particular group uses hashtags. In addition, topics discussed in non-hashtag containing messages were not included. Current research efforts therefore also include the manual classification of a fairly large, random selection of non-hashtag containing twitter messages. The second reason is that not every Dutch citizen is active on Twitter. Because very little information is available on Twitter users, it is anticipated that it will be difficult to relate Twitter findings to the Dutch population as a whole. We are therefore also looking into ways to obtain additional information on the users of Twitter in the Netherlands.

5. Global positioning system data and traffic loop information

Statistics Netherlands publishes quantified traffic data among others to support policy makers. Currently this data is predominantly collected from transportation companies by questionnaires with a relatively high response burden. For vehicles equipped with a route planner, GPS-data could provide very detailed information on their whereabouts (Ma et al., 2010). A marked, practical advantage of GPS-data is that transport and traffic data are captured automatically at highly frequent rates yielding instantaneous, real-time, and accurate information, including position, direction, space-mean speed, and time. This not only could be a good alternative for the route information requested from transportation companies, it could also be used to provide real-time traffic information and provide valuable information on the downside of traffic, such as air pollution, noise, and accidents.

5.1 Data collection

The last decade has shown a massive increase of traffic state signals generated by GPS.

The latter is a space-based global navigation satellite system, consisting of 24 to 32 satellites, an extensive control system and many users. For traffic, GPS provides location data (latitude, longitude, and elevation) in time intervals of approximately one second, together with speed and the number of satellites that identify the source. Speed measurement with GPS is based on a series of 'track points' of position estimates at regular time intervals. The accuracy of the obtained vehicle speed is very high, although it varies with the number of tracked satellites and their geometrical distribution above the horizon. If the number of measured satellites is three or four, then the observations are generally regarded as reliable. However, the adoption of GPS-data is seriously hampered by the fact that still relatively few vehicles are equipped with it. In addition, our own requests for GPS-data revealed that not all transport companies are eager to provide this data. Because of these issues, the collected GPS-data will only cover a limited part of all transport and traffic activities in delineated areas and specific time slots. Combining the GPS-data with loop-detector information might solve these issues (Ma et al, 2010).
Loop detectors are the most popular approach to capture traffic data on roads. Usually, loop detectors are embedded in the pavement. When a vehicle passes over the loop or stops within the loop, this changes the current through the loop which is observed by a sensor. Based on these observations, information about traffic flows is generated, such as vehicle counts per hour and time-mean speed of vehicles passing the loop detectors during a time interval. Combining GPS and traffic loop information could solve the weaknesses in both detection techniques. To enable this, a model was developed.

5.2 Discussion

The model of lane-level traffic density estimation is set up for three different cases: the closed lane without on-/off-ramp, the lane with on-ramp or off-ramp installed loop detector, and the lane with on-ramp or off-ramp without loop detector (Ma et al., 2010). In the model the combination of the loop and GPS-data is used to estimate dynamic traffic density, using the advantages of both techniques and taking into account that the number of GPS vehicles is limited. The travel time of a GPS vehicle is an essential concept to get the time boundary of density measurement, which indirectly represents the space mean speed. Later, in order to address the traffic density in a statistical way for official statistics, up-scaling to selected time intervals and road segments is carried out. Both temporal and spatial up-scaling are applied using a statistical weighting strategy (Ma et al., 2010). Current efforts focus on the collection of data to test the application of the model and its outcomes.

6. Conclusions

In this paper four potential new data sources for statistics are introduced and the first results obtained are discussed. From the above it is clear that one data source can be used directly (price information), other require more research (Mobile phone and Twitter), and one has just arrived at the data gathering stage (GPS-data). Regardless of this status, all four data sources identify a similar problem: the selectivity of the data that is or can be collected. For price information on the internet, this is something to be aware of and it will certainly become an important issue when an increasing number of people buy

products online. National Statistical Institutes need to be increasingly aware of this issue and its effect on the (total) economy measured. For the other three data sources, the selectivity of the data is a very serious problem. Not everybody is calling on his/her mobile phone, is active on Twitter, or has a route planner in his/her vehicle. There clearly is a selective group for which this data is available. Regardless of the outcome(s) of our studies on the potential use of these data sources for statistics, we must -in the end- be able to correct for the non-representativity of the data obtained. For GPS, combining the data with loop traffic information and introduction of a model is a possible solution. For mobile phone data and Twitter, auxiliary information of the users needs to be gathered in order to investigate the representativeness of the data for the Dutch population as a whole. It will be a substantial challenge to develop models to correct for the selectivity of the data in these sources.

7. Acknowledgments

The authors are grateful to Frank Harteveld, Merijn van Pelt, Edwin de Jonge, Martijn Tennekes, Bart Buelens, Mark van de Ven, and Joyce Neroni for their valuable contributions to the work described in this paper.

References

Bethlehem, J. (2010) *Statistics without surveys? About the past, present and future of data collection in the Netherlands*, Presentation for the 2010 International Methodology Symposium of Statistics Canada, October 26-29, Ottawa, Canada.

Dialogic et al. (2008) *Go with the dataflow! Analysing the Internet as a data source*, Report for the Ministry of Economic affairs, version May 13th.

Hoekstra, R. et al. (2010) *Automated Data Collection from Web Sources for Official Statistics: First Experiences*, Internal report, June version, Statistics Netherlands.

Ma, Y. et al. (2010) *Estimation of Dynamic Traffic Densities for Official Statistics based on Combined use of GPS and Loop-Detector Data*, Paper for the 90th annual meeting of the Transportation Research Board, 23-27 January 2011, Washington D.C.

Ossen, S.J.L. et al. (2010) *Quality framework for registers applied to online price information and offline route information*, Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

Roos, M. et al. (2009) *Innovative data collection: new sources and opportunities* (in Dutch), Discussion paper 09027, Statistics Netherlands, Heerlen.

Roos, M. et al. (2010) *Using cell Phone data for statistics*, Internal report, December version, Statistics Netherlands, Heerlen.

Snijkers, G. (2009) *Getting Data for (Business) Statistics: What's new? What's next?* Paper for the 2009 European NTTS conference, Brussels, Belgium.