Statistics Netherlands

Memorandum

POC Event driven processing Economic Demography[i]

Johan Lammers

23 August 2013

summary Policy-makers and the public at large want to be informed in a fast way about the development of the society, especially in these times of economic crises. Statistics Netherlands provides daily information. This information refers to very diverse topics. However, the timeliness of this information should improve. Statistical processes are not designed to respond immediately to actual developments. Traditional statistical processing is characterized by estimating a new state (with or without surveys), comparing this estimate with the previous one and arguing whether the differences are caused by events. Altering this concept by starting with the events and, thus, compiling an estimation for the new state offers a range of benefits: more timely statistics, and decreases in response burden, in the complexity of the process and in the effort to compile the statistics.

In a Proof of Concept (POC) it is shown that it is possible to respond to current events in a statistical process and that there is no need to wait for a new state. The events are the daily entry and exit registrations of companies at the Chambers of Commerce. The registration of the Chambers of Commerce is one of the main sources for the General Business Register (GBR). The statistical process concerns the compilation of the economic demography. This statistic provides information on the population of business units (number, structure and dynamics) in the Netherlands. Currently, the statistic is being redesigned from a yearly statistic based on a survey to a yearly/quarterly statistic based on register data, mainly on the GBR. The POC concerns a daily statistic

.

The estimation method that is developed within this POC is driven by events and is based on historical input and output data of the GBR. In the development of this estimation method the operation of the GBR is considered as a black box.

Comparison of the results from the estimation method with the GBR output shows that it is possible on the basis of daily events to estimate the population of business units.
Event driven statistical processing is a concept with large potential to improve the compilation of official statistics.

keywords Event processing, Economic Demography, Proof of concept

1. Event processing[ii]

Event processing is a part of event driven architecture.

Event-driven architecture (EDA) is a software architecture pattern promoting the production, detection, consumption of, and reaction to events.

An *event* can be defined as "a significant change in state". For example, when a consumer purchases a car, the car's state changes from "for sale" to "sold". A car dealer's system architecture may treat this state change as an event whose occurrence can be made known to other applications within the architecture. From a formal perspective, what is produced, published, propagated, detected or consumed is a (typically asynchronous) message called the event notification, and not the event itself, which is the state change that triggered the message emission. Events do not travel, they just occur.

Event processing is a method of tracking and analyzing (processing) streams of information (data) about things that happen (events), and deriving a conclusion from them.

There are three general styles of event processing: simple, stream, and complex. The three styles are often used together in a mature event-driven architecture.

Simple event processing concerns events that are directly related to specific, measurable changes of condition. In simple event processing, an event is notable if it initiates downstream action(s). Simple event processing is commonly used to drive the real-time flow of work, thereby reducing lag time and cost. For example, simple events can be created by a sensor detecting changes in tire pressures or ambient temperature.

Events (for instance orders, RFID transmissions) are screened for notability and streamed to information subscribers. Stream event processing is commonly used to drive the real-time flow of information in and around the enterprise, which enables in-time decision making.

Complex event processing (CEP) allows patterns of events to be considered to infer that a complex event has occurred. Complex event processing evaluates a confluence of events and then takes action. The events may cross event types and occur over a long period of time. The event correlation may be causal, temporal, or spatial. CEP requires the employment of sophisticated event interpreters, event pattern definition and matching, and correlation techniques. CEP is commonly used to detect and respond to business anomalies, threats, and opportunities.

These events may be happening across the various layers of an organization as sales leads, orders or customer service calls. Or, they may be news items, text messages, social media posts, stock market feeds, traffic reports, weather reports, or other kinds of data. An event may also be defined as a "change of state," when a measurement exceeds a predefined threshold of time, temperature, or other value. Analysts suggest that CEP will give organizations a new way to analyze patterns in real-time, and help the business side communicate better with IT and service departments.

An example to illustrate CEP is the treatment of complaints from customers by an enterprise. The enterprise can choose to react immediately or to wait until a certain amount of comparable complaints are fired. In the first case the event processing is simple, in the second case it is complex: you have to wait until enough comparable complaints are fired.

Of course, rarely does the application of a new technology exist in isolation. A natural fit for CEP has been with business process management, or BPM. BPM very much focuses on end-to-end business processes, in order to continuously optimize and align for its operational environment. The integration of CEP and BPM must exist at two levels, both at the business awareness level (users must

understand the potential holistic benefits of their individual processes) and also at the technological level (there needs to be a method by which CEP can interact with BPM implementation).

The financial services industry was an early adopter of CEP technology. Today, a wide variety of financial applications use CEP, including profit, loss, and risk management systems, order and liquidity analysis, quantitative trading and signal generation systems, and others.

Time series data provides a historical context to the analysis typically associated with complex event processing.

The ideal case for CEP analysis is to view historical time series and real-time streaming data as a single time continuum. What happened yesterday, last week or last month is simply an extension of what is occurring today and what may occur in the future. An example may involve comparing current market volumes to historic volumes, prices and volatility for trade execution logic. Or the need to act upon live market prices may involve comparisons to benchmarks that include sector and index movements, whose intra-day and historic trends gauge volatility and smooth outliers.

2. Concept

For this POC we choose to apply event processing within the statistical process to compile economic demography. This statistic has a close relationship towards available events. Also, time series are available to support event analysis and to use in the resulting processes if appropriate.

This statistic is in redesign from a design based on yearly surveys towards a design based on the use of the General Business Register (GBR). In the POC the focus is on the latter design. The identification, shortlisting and selection of relevant events is based upon the GBR as the main source for economic demography. The same applies for the algorithms to transform the event information towards estimations.

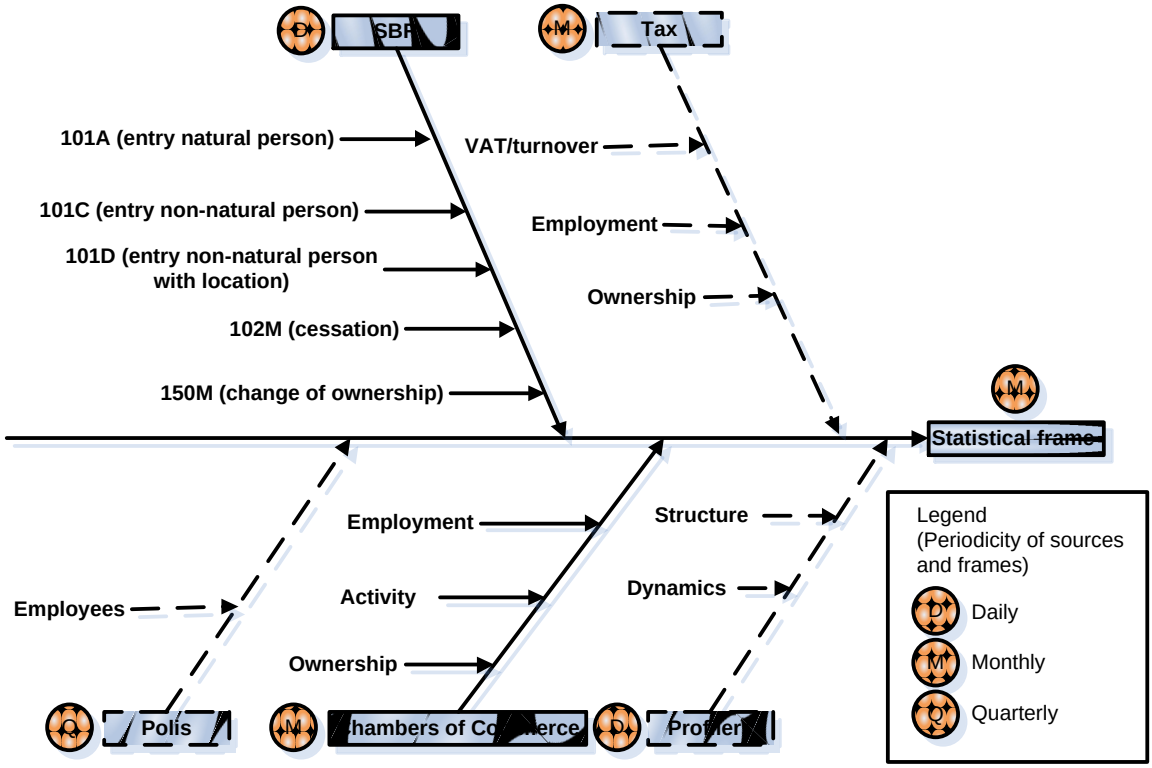
The GBR in the Netherlands uses several sources, indicated in scheme 1.

The Single Business Register (SBR) is a, government-wide, registration of enterprises. After January 1 2014, this will be integrated with the register of the Chambers of Commerce. Polis is a co-operation between the tax-office and the social security boards for the registration of jobs, wages en social benefits. The statistical frame is the basis to compile the statistics.

Scheme 1 denotes the total gross list of inputs. The information for all these inputs is collected as one or more types of events. This may result in (rather) complex event processing. The differences in periodicity contribute significantly to the complexity and to the quality of the resulting frames.

The solid lines indicate the scope of the proof of concept. Event information produced by SBR, from 2014 integrated with the chambers of Commerce, is input,. The influence from sources with dotted lines on the statistical frame (and indirect on the statistic of economic demography) will be neglected.

Scheme 1. Sourceinformation for the GBR



3. The proof of concept

Event processing promises to deliver many benefits:

- The speed of straight through processing (immediate processing) directly in succession of the collection of data will improve the timeliness.
- The spread of information over time is distributed almost uniformly. Peaks and pits in the capacities needed to process information can be avoided.
- Events affect various sources. Event wise processing can be performed on all these sources simultaneous. This avoids synchronisation problems (for instance: matching problems).
- The design of the process will be more straightforward and maintainable.

Within the POC the main goal was to show the improvement in timeliness of the statistic, without substantial loss in quality. For this purpose the complexity of the event process is reduced. The scope was limited to:

- the use of the events delivered by SBR (in scheme 1 identified by 101A, 101C, 101D, 102M and 150M in scheme 1). These events are notable for the births and cessations of enterprises. Within this set the contribution of 101A, 101D and 102M is quite dominant.
- the statistic on the number of enterprises and on the dynamics of this number.

The consequence of the scope reduction is that the effects from the other events are neglected. The impact of these effects is estimated.

4. Analysis

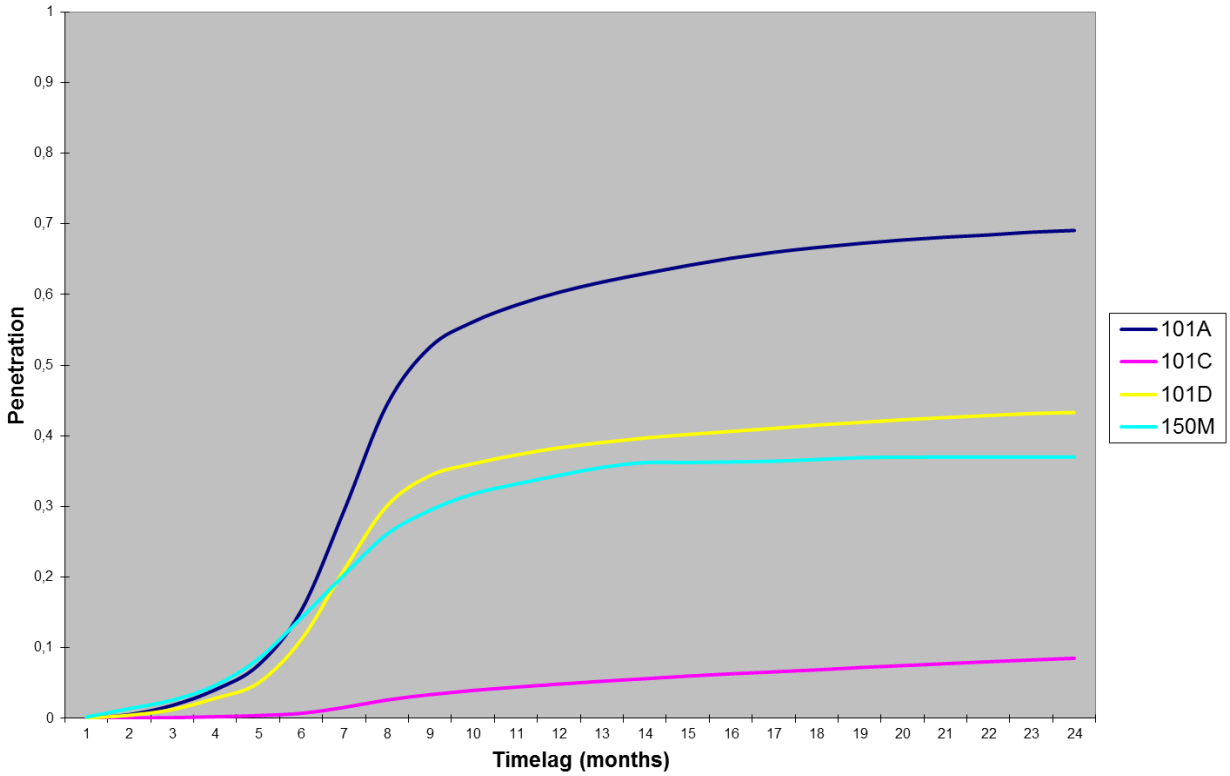
For the analysis to prepare the event processing, we used:

- the time series data of all the received events from SBR, starting in 2006
- the time series data of all the statistical frames, starting in January 2007
- knowledge about processing of frames, with special attention to incidental, periodical and structural anomalies

The information was analysed as cohorts, determined by the type of the event, the year and the month of registration by SBR. So we used the cohort of registered new entries of natural persons within the SBR (101A) in January 2006, a cohort of 101A in February 2006 etcetera.

For every element of a cohort can be analysed if, how and when it effects the statistical frame: the penetration of a cohort. When a new entry leads to the birth of an enterprise in the statistical frame, it has penetrated the frame. Graph 1 shows an example of the penetration as new entry in the frame, per type of event, as an average over all the cohorts.

Graph 1. Average penetration as new entry in the statistical frame per type of event



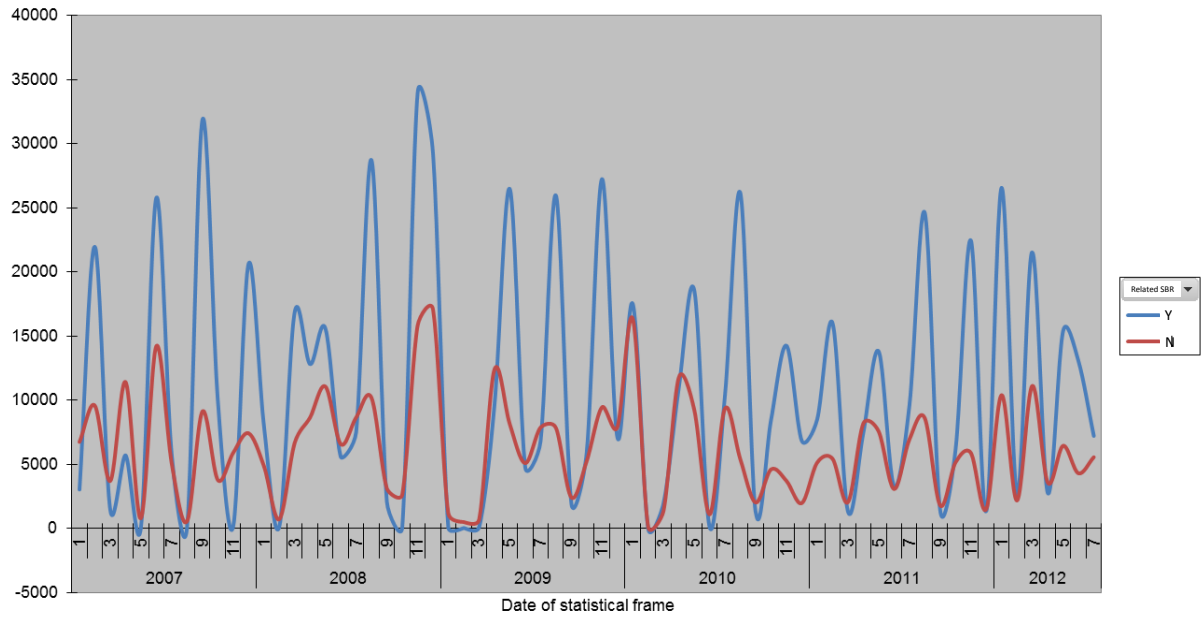
Graph 1 shows the differences in effects on the frame by type of event. The reactions differ by type of event.

In relation to the goal of the POC the graphs shows that daily real-time estimates of the number of enterprises can be produced. For the new entries it is feasible to predict/estimate the number in the future.

Furthermore, the graph shows that the about 50% of the expected penetration is is realised between 6 and 7 months. The time lag is composed of real, economic, and into nonreal, registrative, components, The real component is the time to actually start up the economic activities by the enterprise, after the registration in SBR. The nonreal component is the time needed to collect and process the information. Unfortunately, we could not estimate directly the contributions of these components. A first impression of the contribution of the nonreal component can be derived from the periodicities of the sources and the frames: Quarterly data from 'Polis' causes an average delay of 1.5 months, monthly frames of 0.5 months etc. Adding up all the sequential effects tends to an estimation of 4 months for the nonreal component.

Analysing the effects of the cohorts is one side of the medal. The other side starts with the new entries in the frame. Are they in fact related with events from SBR?

Graph2. New entries in the frame, with/without relation to SBR



Graph 2 indicates that the majority of the new entries is related to the SBR, and that a remarkable part is not correlated. The relation is determined as a direct follow-up relationship between a new entry in SBR and a new entry in the frame within a period of 24 months. The estimation of the new entries without correlation to SBR is not included in the estimation model of the POC. Balanced against the cessations without correlation the number appears steady/constant, which indicates that the dynamics of the number of enterprises highly depends on the correlated information from the SBR.

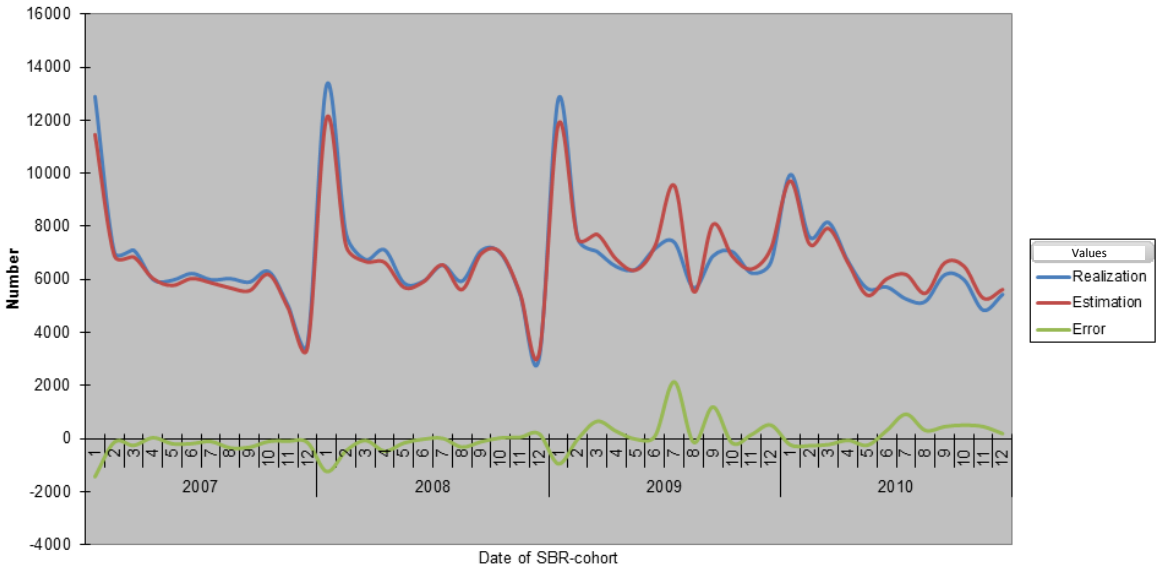
The graph shows a strong pattern of temporal peaks and pits. This pattern is not caused by the SBR! The main cause for this pattern is the processing of the 'Polis'-data. The graphs also shows some incidental deviations (January 2009, January 2010). In these month a major change is implemented in the processing of the GBR. For the purpose of the POC (event processing of the statistic economic demography), it is important to determine how to deal with the incidental and temporal deviations. They are nonreal by nature. The estimation model to process the SBR events can include or exclude these anomalies.

Finally, there are some structural changes. They are not very clear from the graph. In January 2012 and January 2013 the GBR changed the processing of new entries with the aim to reduce the nonreal time lag. Effectively this reduced the time lag with 2 months, on average. For the estimation model, again, the question arises to include or exclude this effect. In the POC, the model was changed.

5. The results

Using the analysis of the cohorts and the frame changes, an estimation model was implemented. The average penetration has been used to initiate the model. For every incoming event from the SBR, the probability to penetrate as a new entry or as a cessation within the frame is used, including the time lags. This relatively simple model results in an estimation of the new entries, the cessations and the total number of enterprises in the Netherlands. The quality of the estimates is illustrated in the next graph (estimating the new entries, correlated to the SBR).

Graph 3. Realization and estimation of new entries, related to SBR



Graph 3 illustrates the quality of the estimates. It shows the results from the model, related the de dates of the cohorts. All the new entries in the SBR from July 2009 will finally result into 8000 enterprises in the frame. The estimation model results in 10000, an overestimation of 2000 enterprises. The graph shows that the average penetration rate results in a fair estimation. The estimation error can be used to analyse the differences and to improve the model. It also indicates anomalies in the source data and in the processing of the GBR. In July 2009 a lot of registrations (without any economical contributions) are entered in the SBR, resulting in the overestimation of 2000 enterprises.

6. Conclusions

This POC proofs the feasibility to use events in the processing of statistics. It is possible to improve the timeliness dramatically. The quality can be improved by fitting the estimation model and/or by adding new types of events. More details can be worked out by using additional information on the economic activity and the size of the enterprises.

The model can also be used to trace anomalies in the collected data and in the processing of the data.

Finally the model can be used to estimate changes caused by redesigns.

i With special thanks to Frank Aelen, Martin Luppes, Magda Slootbeek-van Laar and Dick Woensdregt for reviewing this paper.

ii This chapter is mainly based upon Wikipedia(event driven architecture and event processing)