αY AnalyticalResearch
Quantitative & Qualitative Quality Research

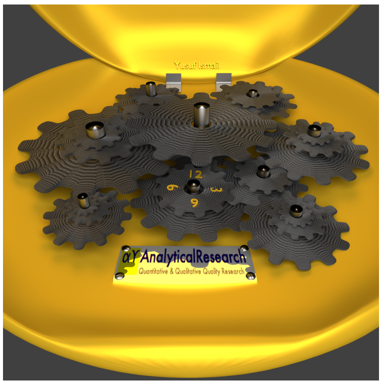
December 2013, No 1

Research & Insights

Analytical Research Ltd is an independent research company that provides professional quantitative and qualitative research for both the public and private sectors.

We are experienced in analysing large data sets, especially those that are not easily processed using 'off the shelf' statistical software. This includes multi-dispersed data, big data, and routinely collected records.

Analytical Research Ltd is registered in England and Wales, No. 6933835.
Station House, Connaught Road, Surrey, GU24 0ER.



Longitudinal Data Analysis

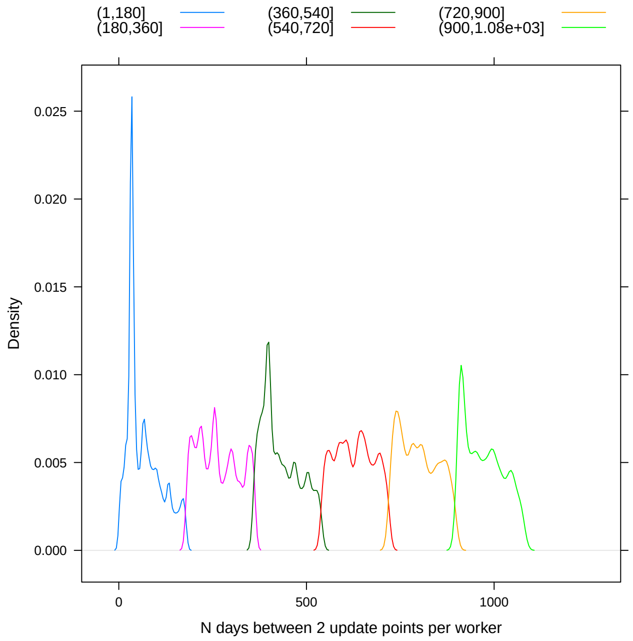
Analysis of longitudinal records has been in use for a long time. However little has been taking place in the social science research area in the UK. Even fewer have been taking place in the workforce sector. There are many reasons why this is happening: lack of longitudinal data; lack of skills required for this kind of analysis; and more importantly lack of appreciation of what kind of information this analysis may bring. No doubt, this kind of analysis is relatively hard, and it gets harder when the size of the data is larger, or sometimes huge (recently referred to as big data). However, the gain from this kind of analysis is substantial compared to multi snapshot data analysis carried out in an attempt to measure change over time. In this article we highlight key factors we encountered when analysing this kind of data.

Longitudinal Records

One of the challenges you would expect when carrying out longitudinal analysis is to find longitudinal records. You would expect to find measurements about a group of subjects collected across a defined time line, however, that is rare in social contexts according to our experience. In a recent project, we needed to carry out a longitudinal analysis over routinely collected workforce data. The sample covered around 700,000 workers, totalling over 9 million records. Each worker has between 1 to 40+ records. These records are spread over 21 standalone data sets with varying structures. The records were collected from different sites and a percentage of these records were redundant, for example, if an employer viewed a worker's record and clicked save/update in the process, it would have entered the data collection process as a new record for that worker. Before carrying out any analysis over such data, we had to go through a lengthy process of cleaning up this data and generating metadata statistics and analysis about the records. Because of the sheer size of the data, we used many visualisation techniques and graphs to find our way through. This also helped deciding what data points to consider for the analysis, which approach to use and many other factors.

Planning The Analysis

We always focus on the research questions/aims. Having a clear idea of what the research aim is about helps during the phase of data cleaning and records' construction. Because we are doing both the analysis and records construction, we focus on the analysis goal and construct the records accordingly, i.e. we don't have a distinction between the two phases. We tested this model of work in several projects, and we found it efficient.



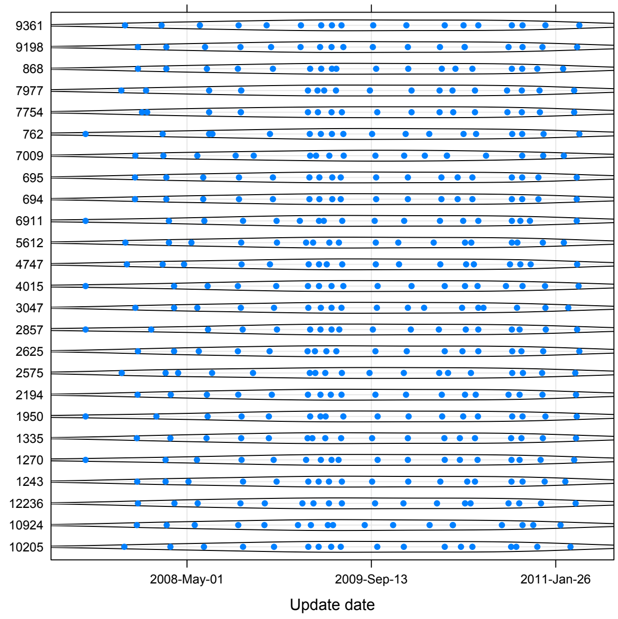
Our Approach

We devised this concept of longitudinal data mapping (which could be thought of, as a proxy to the whole dispersed data sets) that allows us to navigate through the data to extract whatever information is needed, without actually having to deal directly with the data prior to conducting any analysis. It resembles an index of a book, or an actual geographic digital map where you can calculate distances between coordinates without actually having to measure it and scale it up.

There are many benefits as a result of taking this approach. In brief, it allows us to have our metadata analysis, which sets the road for any required longitudinal analysis. Being able to have an overall picture of the data before starting is of a great benefit, as a result, instead of suggesting a research question and proceeding with analysis to find out later there are not enough records to support answering the question, a researcher should be able to look at the metadata first to have a quick and clear indication regarding the feasibility of the research aims. This concept of longitudinal mapping also proved to be very efficient with regards to computer resources utilization. As a result, it provides a scalable solution, powerful enough to sustain the expected large increase in the size of these routinely collected records. The density plot above is an example for information extracted from the map showing the estimated density function for the distance between 2 updates for workers with 2 events. The violin plot is an example for patterns of update periods for subjects with 18+ update events. We check and validate our interpretation from these and other visualisation tools. We discuss our observations about the data with the data provider. Understanding the policy compliance imposed over the employer, help us to understand the data, validate our initial observations and consolidate planning the analysis.

Implementation Behind The Scene

We built this map on a 64-bit multiprocessor Unix workstation with 16 core. We utilised an integrated array of development and analysis tools. C++ has been used to produce machine code highly optimised for vector processing. We used an efficient lightweight database engine that supports parallel computations, and relied heavily on the Unix architecture for the flow of information between all the components including the statistical programming environment. We had to implement several software components that work together to build the map and carry out the analysis. The components and the algorithms were designed and implemented to utilise the highly available processing power. The algorithms were subjected to further optimisation to increase the efficiency of accessing the data by taking advantage of pre-knowledge about its potential usage.



The end result is a system that responds in seconds rather than days, in fact, in our early attempts, a query used to run for 3 hours, however, with our most recent implementation it only takes 10 seconds. We still have queries that may take up to few minutes. That is not bad, especially when considering that other approaches apparently common in use, take over a week to run - yes, it is true.

We hope you enjoyed reading this release and see you soon!