USING CREDIT AND DEBIT CARD PURCHASE TRANSACTION DATA FOR RETAIL SALES STATISTICS

Using Point of Sale Data to Measure Consumer Spending: The Case of Moneris Solutions

Lorant Szabo

Moneris Solutions, 3300 Bloor Street West, 10th Fl, West Tower, Toronto, M8X2X, Canada

Keywords: Data Warehouse, Data Mart, Credit Card Data, Debit Card Data, Point of Sale Data, POS Data, Consumer Behaviour Statistics, Consumer Spending, Retail Sales Statistics.

Abstract: This paper presents how Moneris Solutions stores the credit and debit card purchase transactions that it is processing for its merchants, and the methodology that was invented to process this data to produce consumer spending statistics. The transactions are extracted from the production systems at the end of each day and loaded into a warehouse. Aggregations are executed with pre-established frequency, or on an ad-hoc basis, for various merchant samples to calculate sales growth rates in multiple segments. The results are then matched against known events (e.g. Olympic Games) that may have impacted consumer spending during the analysed timeframe. Alternatively, the results may be presented to measure spending growth between any given two time periods in a specific geographical location or industry.

1 INTRODUCTION

Moneris Solutions is one of the largest credit and debit card processing companies in North America, processing in excess of 2.5bn transactions a year for more then 350,000 merchant locations in Canada and the USA.

The transactions are processed by robust software systems and applications hosted on mainframe systems and non-stop servers. Every day the processed transactions are extracted from the production environment and loaded into a data warehouse. A number of reporting applications and analytical data marts are fed with data from this warehouse on a daily basis.

As Moneris Solutions has the highest market share in Canada, both overall in the country, as well as in most of the geographical and industry segments, we concluded that this data is a good representative sample within most segments – a segment for the purpose of this exercise being the cross-section of a geographical area with an industry (for e.g. restaurants in a given city). Therefore, it became obvious that this data lends itself perfectly to measuring consumer spending on overall aggregate level, or within the desired segments.

A major benefit of using the POS data for this purpose vs. the traditional methods is that the data becomes available shortly after the purchases are completed, typically within a day or two, and it is at the lowest level of granularity, i.e. on transaction level, compared the weeks or months required to collect the research data need for similar statistics.

The drawback is that we have to scan, filter and aggregate millions of transactions, which requires more processing power and strong technical skills compared to the traditional methods.

Due to the proprietary nature of the topic discussed, the paper provides only a high-level, generic presentation of the solution that Moneris Solutions created and implemented to store and process this data and to deliver statistically and economically relevant numbers, reports and information. Nevertheless, most of the ideas presented can be implemented in similar situations, assuming appropriate analysis is performed to adopt these ideas to the peculiarities of the specific transaction processing systems and POS data structure.

Szabo L. (2010).
USING CREDIT AND DEBIT CARD PURCHASE TRANSACTION DATA FOR RETAIL SALES STATISTICS - Using Point of Sale Data to Measure Consumer Spending: The Case of Moneris Solutions.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 447-452
DOI: 10.5220/0003030104470452

2 MONERIS BACKGROUND

The Moneris Solutions data warehouse where the daily transactions are stored holds in excess of rolling 13 months worth of transaction level data, including all the relevant attributes of the transactions, such as date/time, $ amount, transaction type (purchase, refund), store id, store name, store location, store industry, card type (credit, debit), card origin country, etc.

With all the different types of transactions the warehouse holds more than 3bn transactions. Millions of transactions are loaded into the warehouse every day, while old transactions are deleted to make space for the new data.

Various stored procedures are scheduled to run during the day to filter, aggregate and transfer the data in the data repositories of various applications (reporting applications, analytical data marts, etc.).

Data can be aggregated based on pre-established rules for consumer spending statistics and reports on a regular basis, and ad-hoc queries and analyses are completed upon request to measure the impact of specific events (e.g. Vancouver 2010 Olympics).

3 THE MONERIS APPROACH

3.1 Sampling Methodology

For any given two comparable time periods a subset, or sample, of merchants is selected whose sales volumes will be included in the comparison to calculate growth rates.

The merchants are first classified into three categories: “Must-include”, “Include-some” and “Ignore”. This is consistent with the sampling methodology commonly used for consumer spending research.

“Must-include” are those merchants that due to their size and/or importance can not be left out from the sample (for e.g. largest department store chain or gas station chain in a country). Typically, there are a small number of merchants within a country that fall within this category and they are well known.

“Include-some” is the large population of merchants that are smaller in terms of sales volume and comparable, or a homogenous group, from the point of view of exposure to market conditions. Therefore it is sufficient to include a subset of them in the sample, as long as their volume weight is not too low compared to the “Must-includes”. In our methodology we decided to include all those merchants from the “Include-some” category that show positive sales volumes exceeding a predefined threshold in both periods. This will result in the exclusion of low-volume merchants, as well as new start-ups and merchants on their way to going out of business (having no or low volume in one period). The implicit assumption here was that the impact of the two would off-set each other. As far as the low-volume merchants are concerned, they would inflate the merchant count in the sample without having a significant volume contribution. In our methodology, actually these excluded merchants are considered to be the “Ignore” group of merchants.

Other approaches can be adopted to define the group of “Ignore” merchants, as well as to select the merchants to be included in the sample from the “Include-some” category. The size and unique mix (bias towards certain segments, types of merchants, etc.) of the market share of the company providing the POS data will determine the boundaries of the viable choices of approaches.

As part of our methodology, a validation of the resulting samples for the segments (cross-section of a geographical area with an industry) is performed to ensure we have a sufficient number of merchants within each segment. Segments that do not meet pre-established criteria are excluded.

The validation is performed automatically for our regular reporting and it may include a one-time analysis in the case of ad-hoc requests.

All of our sampling and validation was implemented in SQL scripts, executed as part of stored procedures on a regular basis, or as ad-hoc queries with slight modifications to meet the requirements of the one-time analysis at hand.

The business logic and criteria embedded in the SQL scripts is periodically reviewed and adjusted as required by the changes in the business environment.

3.2 Time Periods

The changes in consumer spending can be measured for any given time period that captures the effect of one-time events or matches a typical reporting period (calendar month, fiscal period).

The time period measured can be as short as a couple of minutes/hours, it can be specific days, weeks or months, as long as the time periods are comparable from the point of view of the analysis, and the comparison has economic meaning.

The one-time impacts of unique events may impact one, or both of the time periods (e.g. one time sporting event or holyday weekend vs. the weekend before or holiday this year vs. same holiday last year). Whenever the purpose of our analysis is to measure the impact of one-time events, we would include the days of the event in the base period and find an average-looking same-length period in the past. For the purpose of trending we may do the opposite by trying to 'normalize' for the effect of the one-time impact. An example of this would be the exclusion of the days affected by an event from the base period, and the corresponding days from the period we compare against.

The business requirements for the time periods to be compared are defined in consultation with the users, both for the regular reports and for the ad-hoc reporting. These requirements are implemented through date and time variables and 'where' clauses in our SQL scripts. The date/time values are automatically assigned in our regular reporting and are manually changed in the ad-hoc queries.

3.3 Timing of the Data

One of the major challenges we faced was related to the timing peculiarities of our POS data. While the transactions all have a transaction date and time attribute, the transactions sourced from the different type of systems will become available in the warehouse with a considerable time lag after the transaction was completed. In our case this time lag is anywhere between one day to a couple of days.

The length of the time lag varies depending of the type of the systems they are sourced from and depending on the different days of the week. A thorough analysis was performed to establish the optimal timing of the data extracts that offers the best trade-off between delivering result quickly vs. having a sufficient sample of transactions and merchants included in the extract.

3.4 Growth Rate Calculation

As soon as the merchant sample is defined and the timing of the data extracts has been established the actual sales volume data (the purchase transactions completed within the desired timeframe) can be extracted for the merchants in question, and loaded in temporary tables for validation and aggregation.

A validation is typically performed to isolate any merchants for whom we observe excessively high or low growth rates, and an analysis may be performed to decide if any of these merchants should be excluded. For automated reporting the SQL scripts may include business logic that will perform the exclusions based on pre-established thresholds.

The sales volume data from the temporary tables is then aggregated grouping by the various segments that are required for the analysis or report to be delivered (by various levels of geography and industry – e.g. fast food restaurants near the railway station in the capital).

The aggregate sales volume can be defined as the sum total of the purchase transactions that were completed at the merchants included in the sample during the established time period and available at the point in time when the data extraction is executed.

The aggregate sales volume numbers are then divided to obtain the growth rate on each level of aggregation.

To illustrate the calculation, let us consider the following:

SV0 – sales volume for the base period
SV1 – sales volume for current period for which we measure the change in consumer spending
GR – growth rate

The growth rate can be calculated as:

$$\mathbf{GR = (SV1/SV0) -1} \qquad (1)$$

The growth rate is calculated for each aggregate level, for all the subtotals adding up to the total (e.g. Cities within the Provinces then Province, Shoe Stores within Retail, then Retail) and loaded into a final table for reporting purposes.

3.5 Inference

The underlying assumption in our methodology is that the subset of merchants selected for each segment is statistically representative for the respective segments, and we can infer with a high level of confidence and with an acceptable margin of error that the consumer spending trends observed for the sample are similar to those of the entire population within the segment.

This assumption was proven to be right by the fact that the historical sales trends produced using our data and methodology matched closely the trends available from other publicly available sources.

3.6 Validation

We validate our results by various means. First of all, we compared our long term trends against the publicly available consumer spending statistics (Statistics Canada, retail associations, professional publications, etc.) to calibrate and re-calibrate our methodology and the business logic implemented in our SQL scripts. Secondly, after each analysis we conduct thorough testing of the execution of our scripts and we analyse and test in detail a few merchants to ensure the growth rates are reasonable. Furthermore, we assess the reasonability of the overall findings based on our professional experience and industry knowledge. Any trends or numbers that seem to be significantly different from our expectations are subject to further analysis to determine if we have sufficient reasons to trust the results to be reflecting the market forces, and are not caused by anomalies in the data or programming issues.

Parts of our validation process were automated, however, most of it requires human intervention and attention, and is performed by highly skilled and experienced analyst.

3.7 Dissemination of Results

The results are delivered using various media and formats to different audiences. Our numbers were packaged as summaries in merchant facing web-based reports, subscription based printed and web-based detailed reports for merchants, Excel reports for internal users, data summaries transferred into data marts used for internal analysis, and last but not least, are used for media briefings for printed press, news agencies and interviews with commercial TV channels.

4 COMPARATIVE EVALUATION

In this paragraph we will to provide a high-level comparison of the costs involved in producing retail sales statistics via the traditional methods using survey data vs. the Moneris model using POS data.

As Moneris Solutions Corp. is a privately held company we do have strict confidentiality policies in place that present limitations regarding the cost data that we are allowed to disclose in this study. Also, we do not have access to the cost structure and data of the organisations and companies that produce similar retail sales statistics using survey-data.

Nevertheless, as we will show in the coming paragraphs, the difference in cost structure is so obvious and order of magnitude so significant that it leaves no doubt about the cost advantage of the Moneris POS model. Estimates can be easily performed based on country specific salary ranges to quantify the cost advantage in monetary terms.

From Moneris' point of view - or for a matter of fact any other similar card processing organisation's point of view - the extraction and storage of transaction level POS data is a must for a variety of business reasons, such as fraud detection, business performance reporting, cost monitoring, audit, charge back processing, etc.

Actually Moneris started collecting this data long before we started using it for retail sales statistics. Therefore, the costs associated to the extraction and storage of the POS data itself are not occurred as a result of the existence of the retail sales statistics.

The costs directly associated with the production and validation of the retail sales statistics in a POS model similar to the one presented in this paper can be estimated as follows: ~50% time of one database administrator FTE (full time employee), ~50% time of one data analyst FTE and ~50% time of one business analyst FTE, with corresponding management overhead. Based on our experience these are conservative estimates.

An initial 3-4 month one-time development effort for 2-3 FTEs is required to build out the analytical infrastructure (aggregate dataset structure design, databases, SQL scripts, Excel templates, etc.) in addition to the hardware and software costs. This can take more or less depending on the peculiarities of the POS data structure and the technologies used.

In comparison, the survey-based methodology in a market of the size of the Canadian retail market involves collecting and validating data from questionnaires sent out to approximately 10,000 retail establishments.

The success of the data collection is highly dependent on the reduction of non-responses. Data editing at collection is extremely important. Replacement values must be calculated for missing data using different methodologies. This overall process of collection and validation of the data is very labour and resource intensive and for a given month may take in excess of 3-4 weeks to complete. Covering large geographical areas, in the case of larger countries, can significantly add to the cost. The data collected is stored in a warehouse and validated through statistical edits.

Let us assume that from this point on in the production process of the statistics the survey-based model will require the same resources as described above for the Moneris POS model, i.e. one DBA, one data analyst and one business analyst FTE, all at the same 50% rate of utilization.
Moreover, let us assume that the one-time costs associated with building out the analytical infrastructure will be comparable.
According to our estimates these assumptions tend to be on the conservative side to the extent that we believe that the survey based model may require more FTEs or higher utilization rate for the same number of FTEs, as well as more sophisticated statistical software, perhaps with less robust hardware.

Considering the above assumptions, all the costs related to the collection, editing and validation of the questionnaires and the survey data for the 10000 establishments is net incremental cost compared to the costs occurred by using the Moneris POS model and data. While we do not have access to the cost and resource structure of these organisations, it seems reasonable to assume that at least one FTE may be required for each 1,000 establishments, which would imply an additional 10 FTEs required for processing the data for 10,000 merchants compared to our model. The annual cost difference can be estimated based on country specific salary ranges for the additional FTEs.

The above comparison considers only the production cost of the monthly retail sales statistics, i.e. one single product. Using the Moneris model and data we can produce with a marginal incremental costs statistics that compare any arbitrarily defined two time periods (e.g. first two weeks of April vs. first two weeks of December) by changing the date variable values in our scripts and perform parts of the validation process. We could produce a significant number of additional statistics without having to scale up any of our costs and without compromising quality.

Using the traditional survey-based methodology producing the statistics for another timeframe (lets say a two week period) would entail another set of 10,000 questionnaires to be processed, asking for sales volumes for particular days or weeks only, something that retailers may not have readily available. If retailers are willing and able to collect this data at all, this would significantly increase the cost even compared to the collection of the monthly survey data, let alone compared to using the Moneris POS model.

While the above comparison focused on the cost differences only, it is worth briefly mentioning the benefits associated to the marketability of the two products. Using the Moneris POS model we can produce statistics shortly - within 1-2 weeks - after any arbitrarily defined reference period, while with the survey-based model it may take 1-2 months after the end of reference period before the statistics can be made available.

5 CONCLUSIONS

The early result of our research work on this topic indicated that this area is worth being exploited and pointed to the fact that POS data previously not used for this purpose can in fact be successfully used to produce consumer spending statistics.

The thorough analysis and development work that followed the early findings and the successful implementation of our methodology and model, along with the very positive feedback on our results, have proven that this data source and model is an extremely powerful and relatively inexpensive alternative to the previously known ways of producing similar statistics.

At the time Moneris Solutions implemented this model we could not find any research documentation of other models using POS data and to best of our knowledge at the time none of them existed. While this made the early days of our work difficult and challenging, by now other organizations may have successfully implemented similar models.

This paper offers a high-level description of our methodology and more detailed work may follow on various sections of this initiative, each of which may offer a topic on its own for a more detailed paper.

The required data warehouse and database programming implementation is comparable in size and complexity to typical reporting and analytical infrastructure implementations common in larger organisations.

ACKNOWLEDGEMENTS

I would like to extend my thanks to Igor Klepoch, Kevin Tait and Peter Goldsztajn for their contribution, insights and hard work related to this initiative and the subsequent successful implementation of our model.

REFERENCES

As our goal was to produce nation wide retail statistics the research work involved with this initiative was exclusively limited to analysing the publicly available information on the statistics produced by national statistics organizations, with a strong emphasis on similar statistics produced by Statistics Canada. The rest of the work was experimental development (in compliance with the Canadian tax authority's requirement for innovative work). This consisted of iterations of in-house design and development followed by assessment of design alternatives in an attempt to meet the objectives we set for ourselves. Generally known concepts and methodologies were adapted to the unique structure and timeliness of our POS data through trial and error, letting our model evolve until the results indicated that our model actually works.

For the above reasons our references are pointing to documentation publicly available at the websites of these statistical organizations, and not to other research papers.

Statistics Canada, *Retail Trade Survey (Monthly),* http://www.statcan.ca/english/sdds/2406.htm

US Census Bureau, *Advance Monthly Sales for Retail and Food Services,* http://www.census.gov/retail/how_surveys_are_collected.html

US Census Bureau, *Monthly Retail Sales & Inventories,* http:// www.census.gov/mrts/www/mrts.html

US Census Bureau, *Annual Retail Trade Survey,* http://www.census.gov/svsd/www/artstbl.html

Statistics New Zealand, *Surveys and Methods,* http://www.stats.govt.nz/methods_and_services/surveys-and-methods.aspx