Case study for integrating multiple data sources to better estimate product market shares

Nearly all commercial organizations have interest in estimating the market share of their products or services. That is, they find it helpful to know how many customers they actually possess out of all potential customers for their products or services. This is shown in the following equation:

$$\text{Market share for Product X} = \frac{\text{Sales of Product X}}{\sum (\text{Sales of all products in Product X's category})}$$

For small, local companies wishing to accurately measure their market share, the analytic problem is often the lack of data for the denominator in the equation above. That is, they don't know the total number of potential customers in their product market. Larger companies, however, often face the opposite "problem" of having too much data. That is, there are often multiple sources of market information and they often conflict with each other and vary in data quality. For example, industry analysts may report one set of sales or market share estimates for a given company. In addition, the company might have some primary market (survey) research with market share information. And increasingly, big data sources including telemetry data exist and from which a company can compute its market share using millions of transaction-level data points. For example, there are telemetry data sources from which software publishers can estimate the install and usage rates of their own (and competitive) software products. Some of these telemetry data sources collect this information almost automatically. That is, after proper setup, the reporting devices pass along information to a central server without any needed intervention. And these software telemetry data sources reflect populations ranging up to millions of client systems, PCs, smartphones, or servers. But while having information from multiple data sources is almost always better than having less or no information, the many sources of market share information inevitably present clients with a problem question: which data a source is correct? Which of the many estimates of our product's market share should we believe?

The problem

In Central Moment's client work, we find it rare to find one single, ideal data source for market share information on which a company can exclusively rely. Rather, most sources of market information have some positive but also negative characteristics. For example, some data sources are faster and more timely than others. But that speed sometimes comes at the cost of representativeness or accuracy. Or sometimes a data source contains incredible product or customer detail, but then it's also very expensive to acquire and use. We also very often see data sources that report a product's market share at a very high level of aggregation such as over an entire region or at the world-wide level. But the client actually needs estimates at a more detailed country or segment level. In Central Moment's experience, we find that nearly all such sources of market share information can be useful—but only if optimally combined. The client's needs can only be fully met when the multiple sources of market share information are combined in a way that leverages their strengths and offsets their individual weaknesses.

This case study describes one commercial software client faced with three different data sources, each possessing different strengths and weaknesses. The client needed Central Moment's ability to integrate all of them to produce one single, more accurate estimate of its market share.

Data inputs

The three data sources possessed by this client all reported on the market share for one of the client's server products—which we'll call "Product X". One of the three data sources reporting on the market share of Product X came from an original primary market research survey. This survey went out to roughly 500 IT decision makers in each of 10 European countries. Respondents in the survey reported the number of company-owned servers on which product X was installed and the number of servers on which a competitive product was installed. (The nature of this product was such that a server could only contain one version or instance of this product category).

The company's second data source reported the market share of Product X gathered from telemetry data that the client purchased from an outside data vendor. This data vendor estimated the market share possessed by Product X based on millions of servers installed in the same 10 European countries of interest. These servers periodically reported back to the data vendor whether Product X or a competitive product was installed.

The third data source also reflected telemetry data but came from the client company's own internal telemetry reporting system. That is, the client company had begun its own a program of passively collecting software information from a collection of participating companies. But unlike the external telemetry source, this internal data source was biased. This data source could only report which of the client company's own product versions was installed. Still, however, the data source was useful. The source could at least report the market share of Product X relative to the client's other product versions (e.g., Product Y and Product Z). The market share for Product X was simply biased high in this data source since this source's telemetry population excluded competitors.

Table 1 below summarizes the three market share data sources. It describes their different strengths and weaknesses. In sum, none of the three data sources used in this case study was ideal in every way. They each contained specific weaknesses.

Table 1: The strengths and weaknesses of three data sources used for estimating the market share of Product X.

Data source	Strengths	Weaknesses
Primary Research	• Is extremely flexible. One can include other questions of interest when measuring Product X's market share.	• Is relatively expensive and time consuming to collect. • In this case study, the primary research wasn't updated for all countries in all time periods. • Is subject to greater measurement error. Estimates of product market share are subject to a human respondent's knowledge and recollection. Data quality depends on the care and attention given to the survey.

Internal telemetry	• Is unambiguous in reporting; less subject to human error. • Is inexpensive to collect. • Allowed for high-frequency in data collection.	• In this case study, this source did not contain a very large sample of servers and, so, wasn't very reliable. The market share estimates varied greatly each time period. • Was biased. This source omitted servers with truly competitive products and only reported on the client's own products.
External telemetry	• Is unambiguous in reporting; less subject to human error. • Contained extremely large sample sizes in some countries. • Was the most reliable, i.e., it changed the least wave-to-wave.	• Wasn't under client control. The client couldn't influence which countries or segments had a lot vs. a few reporting servers.

For example, the primary research was extremely flexible in terms of the variety of information that it could collect. Along with questions asking what products were installed on the respondent's company servers, the same questionnaire included opinion and attitudinal questions to better understand the background or context around the installed software. Answers to these additional survey questions allowed the client to find the reasons or story behind the market share results. However, since it was a respondent survey, the survey answers were only as good as an IT manager's recollection and attention. This data source also the most expensive. So, the client didn't update the survey and collect new data for every country over each time period of interest. Central Moment accounted for these time gaps in our analysis by allowing for the confidence in this data source to decay over time. This source held less and less influence as the time increased since the previous market share measurement.

As pointed out earlier, the internal telemetry data source was biased. It could adequately report on the market shares possessed by the client company's own products—which included Products Y and Z as well as Product X. But this source failed to contain estimates for the market share held by fully competitive products. The beauty of Central Moment's data integration approach is that we were able to statistically adjust for the omitted competitive products in this source. Central Moment leveraged the strengths of this data source. However, it didn't bias the final market share estimates for Product X.

Finally, while the external telemetry data source didn't suffer from omitted time periods or obvious reporting biases, it suffered from extremely small sample sizes in a few countries of interest. In one particular country, this data source contained just a handful of telemetry-connected reporting servers. Central Moment's Bayesian approach to this project easily accounted for this shortcoming. Our method allows the data sources containing the most accurate information to have the most influence on the final results. In addition, it allows for the countries with weaker data to "borrow strength" from the countries with stronger, more confident data. As a result, the market share estimates for the weaker countries contain less error.

Since each ingredient data source contained different strengths and weaknesses, the client in this case study couldn't rely exclusively on one single data source for its market share estimates. The client

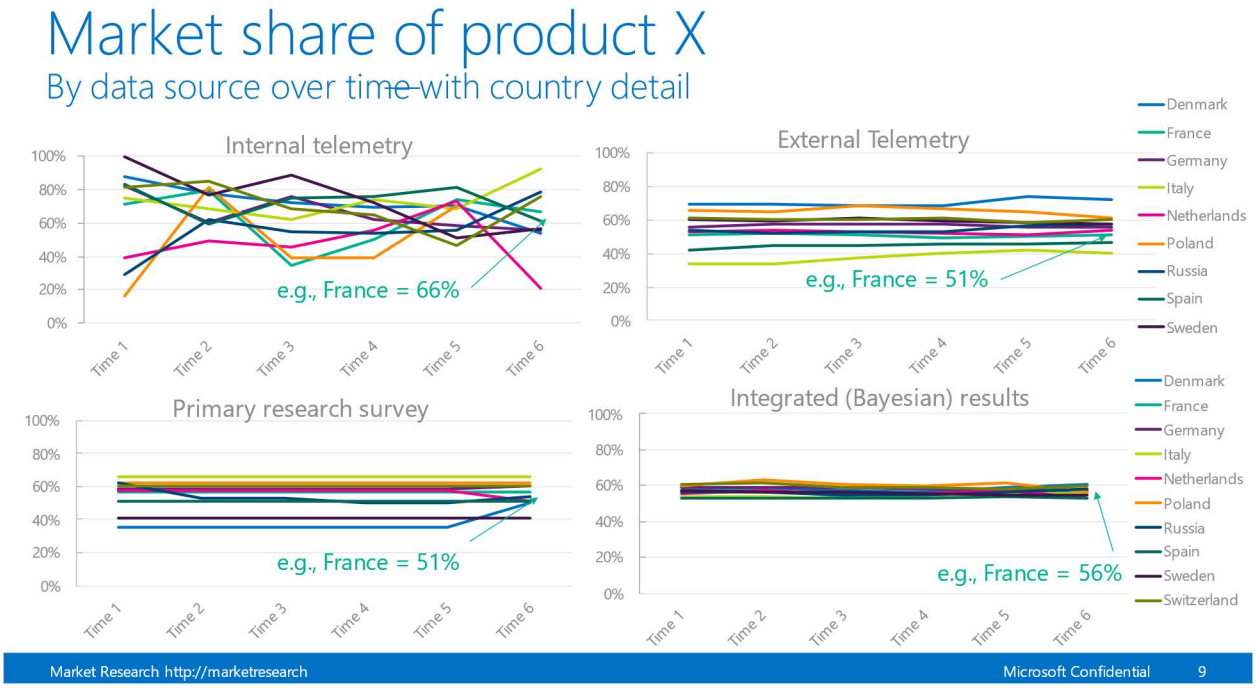
needed an integration of all the different data sources. The client needed Central Moment's approach of optimally combining the data sources to offset their respective weaknesses.

Central Moment maintains such data integration tools as key capability. This short case study can't enumerate our complete methodology. We have other white papers and resources available describing our statistical tools and modeling approach. Most of the time, however, our tools for integrating market share data, such as in this case study, rely on Bayesian statistical methods. Specifically, Central Moment took the data from all three data sources in Table 1 and combined them in a statistical model. The Bayesian element in the model provided the rules and structure for respecting the relative uncertainty of each data source and corrected for known source biases. The end result was a market share estimate that reflected the contribution from each data source in proportion that source's statistical accuracy. This methodology pooled the strengths from the different data sources in a way that produced one single, optimal estimate of Product X's market share—one that was more accurate than any single ingredient data source.

Analysis results

The line graphs in Figure 1 below show the market shares for Product X in 10 countries as reported by the three ingredient data sources over time along with the final integrated (Bayesian) market share result. Again, the internal telemetry source varies the most over time as no country in this source contained a high volume of reporting servers. In general, this source contained the greatest error. By contrast and at the other extreme, the primary market research survey appears to be the most stable or consistent over time in Figure 1. However, please recall that the client company didn't update the primary research every time period. For example, Italy (the country with the highest market share as revealed by the green line in chart #3) was only surveyed once at the beginning of the six time periods and, so, its market share in the chart appears flat over the modeling horizon. That is, for countries not updated with new primary market research, the line chart simply carries forward the market share estimate from the most recent survey (and Central Moment's model assumes less confidence in the carried data points). The external telemetry data source was the one in which the client had the greatest confidence and usually had the narrowest confidence interval.

Figure 1: Market share of Product X over time period as reported in the three different data sources as well as in the final, integrated (Bayesian) result computed by Central Moment (bottom-right chart in the image below). Note that only a few of the market share estimates in the primary research survey (bottom-left chart) were updated during the project horizon. Flat lines in this chart indicate a repeat of the most recently collected market share estimate.



The line chart in the bottom-right of Figure 1 shows Central Moment's final integrated (Bayesian) result. That bottom-right line chart shows the market share estimates after optimally combining and correcting for all uncertainty and known biases in the three data sources. The line patterns in the integrated result reveal three key features of the modeling process. First, the market share of Product X is remarkably smooth over time for each individual European country. The market share doesn't jump around wave-to-wave to the same degree as the input sources. (Again, the overly smooth lines shown for the primary research survey data are a display artifact). This is a typical benefit from this type of analysis. The method reduces wild period-to-period changes when the wild changes in the inputs are likely due to error. The period-to-period change in the inputs that likely reflects true change tends to be preserved in the final result.

The second noticeable feature in final integrated result (in the bottom-right of Figure 1) is that the market share estimates for each country appear to be something of a compromise or average of each country's data inputs. For example, as highlighted in Figure 1, the estimated final market share estimate at Time 6 for Product X in France (56%) is close to the average of its three input values for that period—66%, 51% and 57%, respectively. However, that final share result for France (56%) is a lot more than just a simple average. In fact, the final results for some countries will actually sometimes fall outside the range of their input values! The final result for a country depends not only the country's own particular data but also on the confidence in that country's data vis-a-vis the confidence in the data for other countries. It's not called-out in Figure 1 like it is for France, but the final share result for Poland at Time 5 is 61.6% which is actually below the range of its three inputs (70%, 65%, and 62.3%, respectively).

This "feature" of Bayesian analytics to produce final market share results outside the range of a country's inputs is sometimes disconcerting to company managers. However, such a final result is usually understood after deeper inspection of the data. In this case, Poland suffered from extremely

small sample sizes in its two telemetry inputs at Time 5. So, its relatively high market share estimates in those inputs was highly suspicious. The Bayesian algorithm automatically intervened for Poland and produced a less egregious final estimate. Ignoring these weaknesses in Poland's inputs would have increased the client's risk of error in Poland's estimate and increased likelihood of future disappointment. Indeed, in the very next time period (Time 6) the updated inputs for Poland proved this caution was correct. Poland's inputs at Time 5 proved to be outliers. Poland's inputs for Time 6 retreated to more plausible levels and demonstrated that Central Moment's estimate for Time 5 was likely well advised.

Closely related to the above, the third feature of the final integrated result is that the market share estimates by country are a lot more compressed around a central tendency than seen in any of the inputs. Put differently, the lines in the far-right chart of Figure 1 fall into a much more narrow range than that seen in any of the individual data sources. This is a common result after data integration using Bayesian analysis. The final integrated result pools information from all of the data sources and across all of the countries. Errors in the data sources and by country tend to cancel each other out. As a result, there's less erroneous wave-to-wave variability both across time and across countries. The risk of incorrect outliers in the final result is greatly reduced.

All three of the above "features" of the data integration process are echoed in other, non-commercial applications of data integration. Meta analysis is a broad term used to describe a very similar analytic approach that researchers have applied in numerous other fields of study. For example, cancer researchers sometimes combine the data and results from a large number of separate medical studies to draw more powerful conclusions. If each cancer study is measuring the same effect (e.g., the cancer remission rate) after administration of the same treatment (e.g., an experimental drug), then pooling the studies together usually produces a more accurate final estimate of the drug's effectiveness. Meta-analysis is a closely related cousin of Central Moment's approach.

Conclusion

In this case study, the client went on to use the final data-integrated result for Product X's market share in all of its internal reporting. And Central Moment's analysts added flags or other indicators to alert the client when market shares exceed (or fell short) of expectations. Compared to when the client reported results from just one single source, the new flags and indicators did a much better job of indicating true market share change. Because the new results were more stable over time, they resulted in far fewer false alarms regarding changes in Product X's market share.

Moreover, after observing Central Moment's application of this tool for Product X, the client company adopted it for other server products and for other product groups entirely. The client company ultimately grew to using Central Moment's approach to track and report the market shares for roughly 300 products across more than 50 reporting countries in four different customer market segments. This method of data integration for better market share analysis proved a great success. It even earned the client group responsible an internal company award and financial recognition for its effort.