Proceedings of the 2000 Conference on Information Quality

Using Time Series to Assess Data Quality in Telecommunications Data Warehouses*

(Practice-Oriented Paper)

Renato Busatto
Mobilix Telecommunications
Copenhagen
Denmark
e-mail. rbu@mobilix.dk

20 September 2000

Abstract

The growing complexity of telephone services, particularly in mobile telephony, and its impact upon billing data mean that phone call volume modelling techniques became crucial in the assessment of the accuracy of the information available in telecommunications data warehouses. Time series modelling, normally used for forecasting, provide a suitable tool for this purpose as well. Preliminary experiments show the relevance of the approach in the discovery of missing traffic data.

Keywords. Data Warehousing, Telecommunications Databases, Time Series Modelling, Data Quality Control

1. Introduction

The quality of data in billing systems has for quite some time been an important issue in the telecommunications industry, and has now assumed a new dimension, given the widespread availability of analytical information systems in general, and data warehouses in particular. Increased competition [Mattison 97] means that traffic (i.e. phone call volumes) now has to be monitored closely, and, in the event of sudden traffic reductions, immediate remedial marketing action is required. Given that these measures may have a substantial impact on the company's revenues, it is crucial that recorded traffic matches the real figures as closely as possible. Also, in view of new uses given to data, especially in the context of customer relationship management, more stringent standards as regards the quality of business information became universal.

Pulling in the opposite direction is the fact that the increasing complexity of telephone services, particularly in mobile telephony, leads to the slower gathering of billing data; moreover, the possibility of errors increases alongside service complexity. This means that, besides plain traffic data collection, it became necessary to resort to (usually informal) modelling techniques to generate a more accurate picture of telephone usage based on incomplete information as quickly as possible.

Time series have been consistently used for forecasting for quite some time [Mills 90, Hamilton 94], but their usage in data quality assessment, though certainly existent [Larsen 97], does not seem to be widespread, at least in telecommunications data warehousing. The methods used for forecasting and quality assessment are essentially the same, but there is an important

* Practice Paper. *The 2000 Conference on Information Quality*, MIT, Boston (MA), 20-22 October 2000.

asymmetry. Since data quality assessment is carried out *post factum*, determinant events that could not have been foreseen may be taken into account. For instance, a forecast of the number of phone calls for the first Monday of December 2000, carried out in July 2000, might yield a totally distinct result from the same 'forecast' carried out in January 2001, if at that time it is known that a major snow storm on that particular date disrupted car traffic and led to a substantial increase in the usage of fixed and mobile phones. The introduction of time series modelling in data quality assessment, coupled with special events analysis, can therefore be expected to generate more accurate results than in forecasting *strictu sensu*.

This report examines the possible advantages of the adoption of time series modelling techniques in the assessment of the quality of telephone call traffic data, through a concrete case study. Section 2 defines some of the basic data quality concepts, whereas Section 3 proposes a model for the various cyclic components that make up the phone call volume figures. The case study is described in Section 4. Section 5 proposes refinements for the current model, aiming at increasing its accuracy, and a few conclusions follow in Section 6.

2. Data Quality: Core Notions

Various definitions of *data quality* can be found in the literature – [Tozer 94, Redman 96] are but two sources. For data supposed to represent an aspect of the 'external world' (such as the address field of a customer record, as opposed to a system-generated customer identifier, which has no match outside the database), it could perhaps be said that quality is a relative parameter, in the sense that it has to be measured with respect to a given (external) reference. At least two distinct dimensions of quality can then be taken into account: *correctness* and *completeness*.

The notion of correctness concerns the checking of the adequacy of the representation vis-à-vis the external world. A table is correct w.r.t. an external reality if and only if it accurately portrays the properties of objects of this world. For instance, a *citizen* table is correct w.r.t. the population of a given city if and only if every person named in each record actually lives in this city at the specified address, and was born at the specified birth date, etc.

Completeness involves the total capturing of the external world by the representation. A table is complete w.r.t. an external reality if and only if every object of this world is accurately represented in the table. For the above example, the citizen table would be complete for the given city if and only if every inhabitant was accurately represented by a record in the table.

Clearly, one of the biggest challenges in quality assessment involves reaching the external world, a task that, in the end, amounts to accessing another (independent) representation, against which the data under consideration shall be compared. One of the simplest examples involves customer address data, on the assumption that, besides a company own address records, a (more accurate) database, sponsored by e.g. the social security services, exists. In such cases, the correctness of a company address table can be straightforwardly determined by the direct comparison of the customer name and address fields of both systems.

A more complex scenario occurs when such an independent alternative representation does not exist. This is usually the case when the data describes an activity that is strictly internal to a

company, such as product sales. For these and similar cases, it becomes a necessity to resort to modelling techniques. One of such techniques is examined next.

3. Telephone Traffic Data Analysis

In telecommunications data warehouses, in view of the large amount of available traffic data under the form of *cdrs* (*call detailed records*), a highly granular traffic trend analysis, as well as the construction of models for the assessment of data accuracy, is now possible.

The first question is, which portion of this data should be used for checking the quality of recently collected data. One solution is the use of data that lies in the temporal vicinity of the targeted time period. For instance, if the accuracy of the traffic data of last week's Wednesday has been put into question, it would in principle suffice to examine the figures for the remaining days of last week and resort to interpolation techniques to determine an expected value for Wednesday. However, as mentioned in Section 1, recently collected data as a whole lacks a sufficiently high level of accuracy, hence interpolation turns out to be inadequate for modelling purposes.

Models then should be based on older data, which is much more likely to be complete and correct, even though adjustment parameters might be required. Time series, which are basically functions over discrete temporal domains where measurements of the dependent variable are taken at constant intervals, might be used in this connection. Daily readings of average temperatures on a particular site, or the weekly figures of the total sales of a product are just two examples of time series. Several phenomena that change over time might exhibit both cyclic (or *seasonal*) and non-cyclic (or *trend*) patterns intertwined in a single curve. For analytical and forecasting purposes, it is then relevant to decompose the elements that make up the final figures that describe the events under consideration.

Phone call volumes as a function of time exhibit both linear and cyclic patterns. As in several other cases, seasonal patterns may be identified by visual inspection of diagrams [StatSoft 84]. Such graphical analysis does in principle concentrate on shapes, which reflect relative rather than absolute values. Taking days as the time granule, the following cyclic components can be listed:

- Weekly cycle. Traffic volumes are typically lower in the course of weekends, reach peak values on Mondays and/or Tuesdays, and then recede slightly to (similar) lower values on the remaining days of the week.
- Monthly cycle. The mid of the month may show a slightly lower volume that the first and last days of the period.
- Yearly cycle. Summer holiday months exhibit a clear traffic reduction, whereas, in the winter, an increase might be noticed.

Another factor that affects data quality analysis is the presence of floating core events, which might have a substantial impact on the values assumed by the measurement parameter. This is the case, for instance, of school holidays whose precise dates might vary from year to year. There are also unscheduled events that also affect the parameters, such as public transport system strikes.

Simple pattern descriptions like the above one are actually used by many telephone call traffic data analysts in the course of their routine assessment of the accuracy of the data they are confronted with on a daily basis. The problem with this informal approach, nonetheless, is that it relies on the experience of individual analysts, who will hardly ever formulate it rigorously. Hence, as more (and more complex) distinct components are identified, combining them informally to model a phenomenon is far from a satisfactory solution. This would be clearly the case in a situation where a variety of cycles is identified and informally described, but whose reassembling might become infeasible, especially if rules from distinct cycles pull in opposite directions. The solution therefore is to capture this informal knowledge in a formal model, so that rules can be more precisely applied, and more accurate results, delivered.

A time series model that could help in phone call traffic analysis would have to incorporate all the above described elements. There are a variety of combination techniques, one of which is examined in the coming section.

4. Applications

Time series forecast and modelling encompass various well-developed techniques of distinct degrees of complexity (ARIMA, exponential smoothing, etc) [StatSoft 84]. This case study shall employ the *classical seasonal decomposition* method, which usually takes into consideration four determinant factors: (1) the trend (or linear), (2) the seasonal, (3) the irregular cyclic, and (4) the random, or error, components. The main distinction between the seasonal and irregular cyclic components is that only the former observes a fixed time period (e.g. week, month or semester). When the seasonal component comprises more than one cycle (as in the case of phone call traffic volumes, described in the previous section), the label *multiple seasonality* is usually applied [Mills 90]. The simplified model considered in the present study will not include any errors, but will have instead a *singular events* factor, to cater for unexpected events that reportedly have a noticeable impact upon traffic.

Traffic may be measured from a variety of viewpoints. Just to mention three of them, one might be interested in monitoring the number of calls, call durations, or billed amounts. This study shall be based on the total number of outgoing calls[1]. The following (conditional) equation can then be used to capture the various components described above:

$$t > t_0 \rightarrow V'(t) = V_B\,(1 + D_W(wkd(t)))\,(1 + D_M(day(t)))\,(1 + D_Y(mon(t)))\,L(t) \qquad (*)$$

where t denotes a date (yyyy-mm-dd), t_0, the starting date for forecasting purposes, V', a random variable representing the *expected* total number of phone calls at date t, and V_B, the basic total daily traffic volume. D_W, D_M and D_Y respectively denote the weekly, monthly and yearly cyclic factors of the model, whereas *wkd*, *day* and *mon* are functions that take a date t as argument, and respectively denote the weekday, day of the month and month associated with t. Intuitively, these cyclic factors indicate the increase or decrease of traffic associated for instance with a particular

[1] It is worth mentioning that the figures provided in this section do *not* correspond to the actual traffic volumes for the mentioned period, as recorded in the Mobilix operational data store: changes were made in the basic, cyclic and trend parameters. The actual figures are nonetheless immaterial for the purpose of illustrating the merits of the approach under consideration.

day of the week; from the discussion presented in Section 3, one could expect therefore that $D_W(\texttt{Sunday}) < D_W(\texttt{Monday})$.

L represents the trend factor, which in the present example shall be given by

$$L(t) = (1 + \eta)^{difm(to,\, t)} \qquad (**)$$

η indicates the monthly increase of total traffic, and *difm* determines the difference in full months between two given dates. The final equation for the *actual* total traffic V at date t will nonetheless incorporate two special events components, V_E and w, and is represented as

$$V(t) = V_E(t) + w(t)\, V'(t) \qquad (***)$$

Special events may therefore entirely overrule the effect of both cyclic and trend parameters, depending on the value of the weight parameter $w(t)$. For instance, one might assume that, on national holidays, traffic behaves similarly to Sundays, irrespectively of the weekday on which they fall; in such cases, the weekday component has to be cancelled out.

Figure 1 indicates the values of the cyclic parameters D_W, D_M and D_Y, determined from observations. The mean value of the basic traffic volume (V_B) was estimated as 4.5 million daily phone calls. The values of D_W and D_M were obtained from the analysis of a full month's traffic data, whereas the values for the yearly parameter D_Y came from the direct analysis of monthly summarised traffic figures. A trend increase factor η of 2% was proposed, and the start date, chosen as 1999-01-01.

Once the above parameters are introduced in equation (*), the following expected value can be computed for 1999-12-30 (Thursday):

Basic figure (V_B)	4.5 million
Weekly cycle (D_W)	+60% (x 1.60)
Monthly cycle (D_M)	+12% (x 1.12)
Yearly cycle (D_Y)	+20% (x 1.20)
Trend factor (η)	$+(1.02)^{11}$ (x 1.24)
Expected traffic (V')	11.999 million phone calls

On the assumption that a sharp fall in temperatures occurred on that particular date, a weight factor contribution could be added to the above computation:

Weight factor (w)	+10% (x 1.10)
Extra summand (V_E)	+0.000
Actual traffic (V)	13.199 million phone calls

The usage of the above model in telephone traffic data quality assessment has actually helped not only in uncovering that data was missing in the operational data store (i.e. in pinpointing the *incompleteness* of the representation), but also in estimating the number of missing calls.

Concerning recalibration over time, even though the above parameters are still preliminary and will be refined in the future, experience in telecommunications traffic analysis seems to indicate the relative stability of at least the seasonal components of the model, as they reflect social factors and habits that change slowly, if at all. As for the trend factors, the current expansion of telephone services on a global scale will probably lead to the need of revising them on a regular basis.

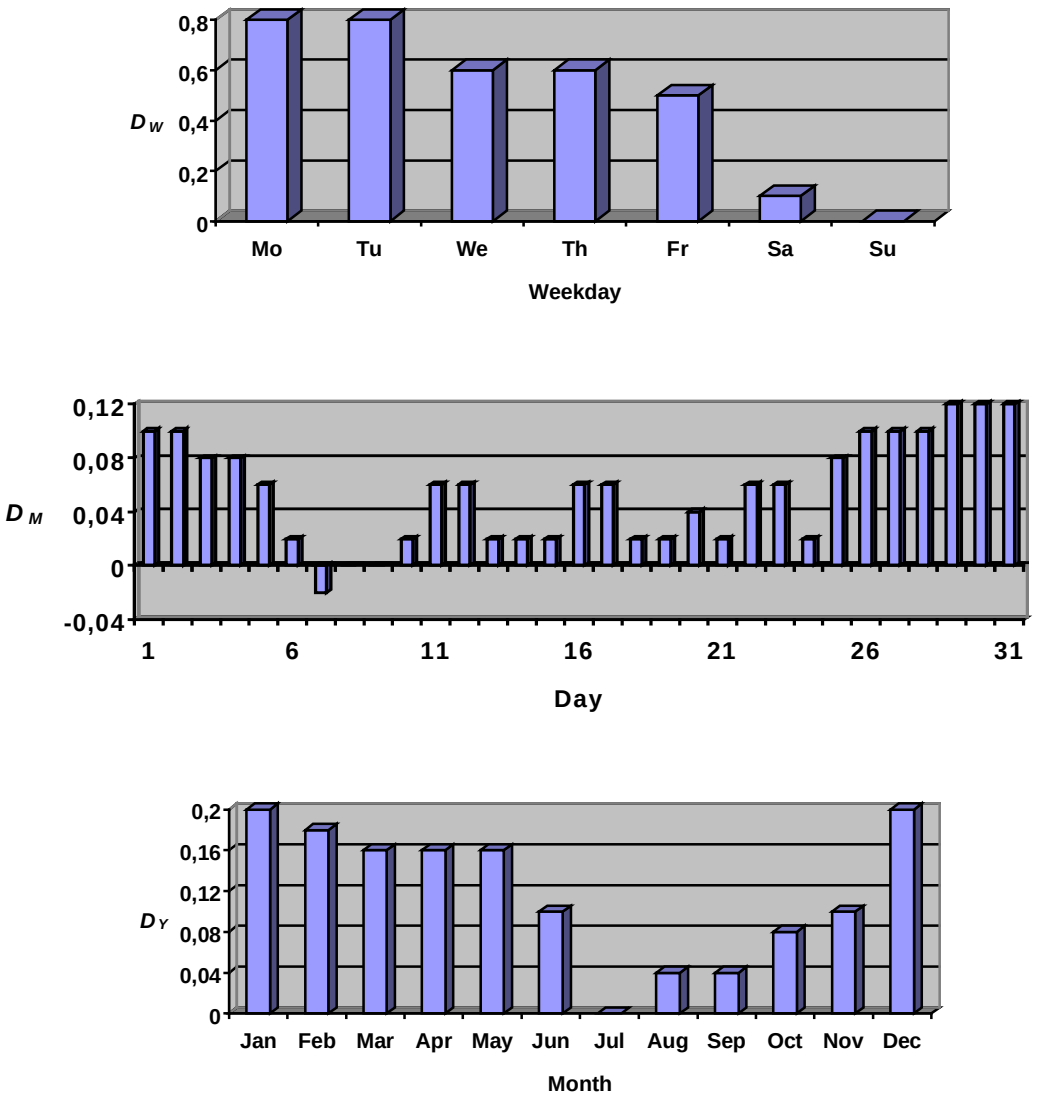

Figure 1. Cyclic Factors

5. Refining the Model

The cyclic component of the above model could be refined even further by the introduction of intermediate cycles, e.g. bimesters, quarters or semesters, and also larger cycles, such as five-year periods or decades. The decision as to whether of not a particular cycle should be introduced depends essentially on evidence suggesting a regular pattern of change in the course of a cycle. For instance, traffic in regions or countries where quinquennial economical plans are enforced might exhibit a recurrent pattern along the duration of longer cycles. A semester cycle, on the other hand, might turn out to be entirely irrelevant in a given context if the measurement parameter does not exhibit meaningful changes along the period, except for those already catered for by the other component cycles (e.g. months).

Moreover, the analysis of months and weeks, for instance, might be carried out either as independent time periods in their own right, in which case the average behaviour of the measurement parameter will be examined along the duration of the given period, or else they may be looked at as component of a larger time period, in which case specific features of each component might surface. For instance, as an autonomous time unit, a month might be seen as a 28/29/30/31 days period, for each of which the average daily value of the chosen parameter along each day will be measured. Alternatively, months might be looked at as component of the year cycle, in which case each of them will receive separate attention, i.e. January months will be considered separately from all the others. Figure 1 would then have to give way to a much larger collection of diagrams, which would include twelve (possibly distinct) monthly charts.

Finally, there is a wealth of time series literature in the telecommunications area (e.g. [BDT 2000]) that has not been covered in this report. Past analyses of cycles in this sector could certainly help in the improvement of the model described in the previous sections. Nonetheless, the level of detailed traffic data presently available in data warehouses opens up the possibility of a traffic segment analysis that was basically infeasible in the past. More accurate models could be obtained if, instead of looking at total daily figures, traffic could be split between the private and business segments. Patterns might differ substantially for these segments, at least along certain cycles: even though the overall traffic falls during weekends, it is usually the case that it increases for the private customer segment on Saturdays and Sundays. As in many other contexts, the more refined the model, the more precise the assessment of data quality. New opportunities for investigation in this area are therefore opened to data analysts and data quality inspectors alike.

6. Conclusions

Data warehousing makes available to data analysts, in many situations for the first time, historic data at a highly granular level. Previous statistical and historic analysis, which would normally have to rely on aggregated and summarised data, can now benefit from the access to virtually all the raw data collected by the operational systems.

As a first step towards exploiting all this information for data quality assessment purposes, time series superposition techniques have been put to the test, and have proven their efficacy in the context of data quality assessment in telecommunication data warehouses, especially when unexpected events are taken into consideration by the model. Similar techniques could be

extended, refined and applied to any other domain where abundant and detailed historic data exists.

References

[BDT 2000] *Yearbook of Statistics (Telecommunication Services 1989 – 1998).* Telecommunication Development Bureau, 2000.

[Hamilton 94] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[Larsen 97] J. C. Larsen. Long Period Changes in the Transport of the Florida Current and its Relationship to Atlantic Sea Surface Temperature and the North Atlantic Oscillation Index. In *Proceedings from a Meeting on Atlantic Climate Variability* (Lamont-Doherty Earth Observatory), available at `http://www.aoml.noaa.gov/phod/acvp/larsen.htm`, 1997.

[Mattison 97] R. Mattison. *Data Warehousing and Data Mining for Telecommunications*. Artech House, Boston, 1997.

[Mills 90] T. C. Mills. *Time series techniques for economists*. Cambridge University Press, Cambridge, 1990.

[Redman 96] T. C Redman. *Data Quality for the Information Age*. Artech House, Boston, 1996.

[StatSoft 84] StatSoft. Time series analysis. Technical report, StatSoft, Inc., available at `http://www.statsoft.com/textbook/sttimser.html`, 1984.

[Tozer 94] G. V. Tozer. *Information Quality Management*. Blackwell, Cambridge (MA), 1994.