Soft Solutions®
information has never been this intelligent

Delivering intelligent business solutions to World Class Retailers

www.ibs-softsolutions.com

Soft Solutions

Ibs AMS

Data Cleaning for sales modeling

Embedded Cleansing engine:

Theory and concepts

Table of Contents

1. Executive summary

1.1. Main benefits

Ibs Analytics is based on a sales forecast engine developed by Soft Solutions to be a strategic tool for the retailer to anticipate consumers' needs. It implements concepts of various scientific fields, like data mining, statistics, signal processing and mathematics, fitted to the business process of retailers.

Moreover, *Ibs Analytics* is an embedded solution with real-time analysis, which provides accurate results with total transparency in simulations and set-ups. Indeed, these results can be obtained, through business-oriented reports, at various consolidation levels (item, segment, category, region, etc…) in easily understandable metrics.

Based on the generated models (trend, seasonality, price elasticity…), *ibs Analytics* performs an average accuracy of ~75% at the item level and over 90% at the category level, for all kind of items including items with actually low turnover.

Furthermore, *Ibs Analytics* has an impact on every level of the retail working chain:

- It strengthens the operational strategy and bring decision-making tools to fulfil objectives
- It avoids being over stocked or out of stock
- It provides an homogenous system to both analyze and predict needs
- It gives consistency to the operational strategy with pricing, assortment and marketing using the same forecast engine
- It induces a quick return on investment

1.2. Objectives

This report aims to:

- Present detailed process of information cleaning
- Provide comprehensible figures and tables to illustrate the goal of each step of the process;

1.3. Contents

The following pages are structured into three sections:

- Overview of the sales forecast concepts
- Step-by-step data cleaning process of *ibs* Analytics
- Conclusions and Reference papers

2. Introduction

2.1. Ibs Analytics embedded within Ibs - Suite

Achieving simulations in the retail business is one of the most strategic parts when defining and applying both marketing and operational policies. Indeed the ability to measure impacts of different decision is the key-part in order to take the right decision.

Forecasting the sales has a central position as it interacts with every phase of the retailer framework (Fig 1). Therefore *Ibs Analytics* is an asset, which brings vulgarized science to business. Indeed, a coherent forecasting strategy among all departments and retail activities will induce a reduction in operational delay and cost due to impact of better forecast and a better visibility among various departments:

Fig. 1. *Sales forecast and the retailer workflow*

Outlier :

In statistics, an outlier is an observation that is numerically distant from the rest of the data. Outlier is historically as:

*"An outlying observation, or outlier, is one that appears to deviate **markedly** from other members of the sample in which it occurs."*

Outlier detection:

Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data.

The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, drawn from the full gamut of computer science and statistics.

- ***Ibs* CATEGORY** and ***Ibs* REPORTING:**

Objectives Management set strategies, which impact the all business process. Ibs Analytics takes into account intelligent distribution of objectives in forecasting and provides a real-time follow-up on achievements.

- ***Ibs* ASSORTMENT:**

Ibs Analytics is also useful in **Assortment** with assortment size optimization and halo & cannibalization measurements. Moreover, it computes benefits simulations in various metrics (units, sales, margins).

- ***Ibs* CENTRAL** and ***Ibs* STORE**:

Operational Processes Optimization implies an ordering optimization in quantities, which depends of sales forecast, costs and rebates …

- ***Ibs* SPACE PLANNING:**

Products Display in the store plan or the store planogram with shelves constraints management is affected by the sales forecasts.

- ***Ibs* PRICING** and ***Ibs* - PROMOTIONS:**

Sales & Marketing Policy Optimization in order to fulfil objectives is strongly connected to *Ibs* Analytics, which respects constraints & controls management and provides business indicators estimation.

Marketing :

Marketing is defined as *"the process of management responsible for identifying, anticipating and satisfying customer requirements profitably."* by the Chartered Institute of Marketing.

In retail, marketing includes advertising and various promotion strategies like:

- Price reduction offer
- Buy one Get one offer
- Bundle offer
- Extra bucks offer

2.2. What is a forecast?

A sale is the result of the consumer perception of several Features (Fig 2) such as:

- The banner strategy, which has a huge impact on the consumer behavior by defining prices, promotion policy and marketing.
- The local microclimate, it is induced by the concentration of competitors and the type of the area (rural or downtown).
- The in-store availability and accessibility are also key features in a sale, with the assortment strategy and the planogram disposal.
- Some external factors are to be taken into account like unemployment, growth rate, inflation rate, which have an effect on the consumer purchase's budget. Moreover, household happiness can increase sales, as people are willing to buy.

Although some of theses factors are hardly measurable, some others are well known and even decided by the retailer. Moreover, a sale can also be defined by some item's specifics: general trend, seasonal cycle, seasonal peaks, price changes, etc.

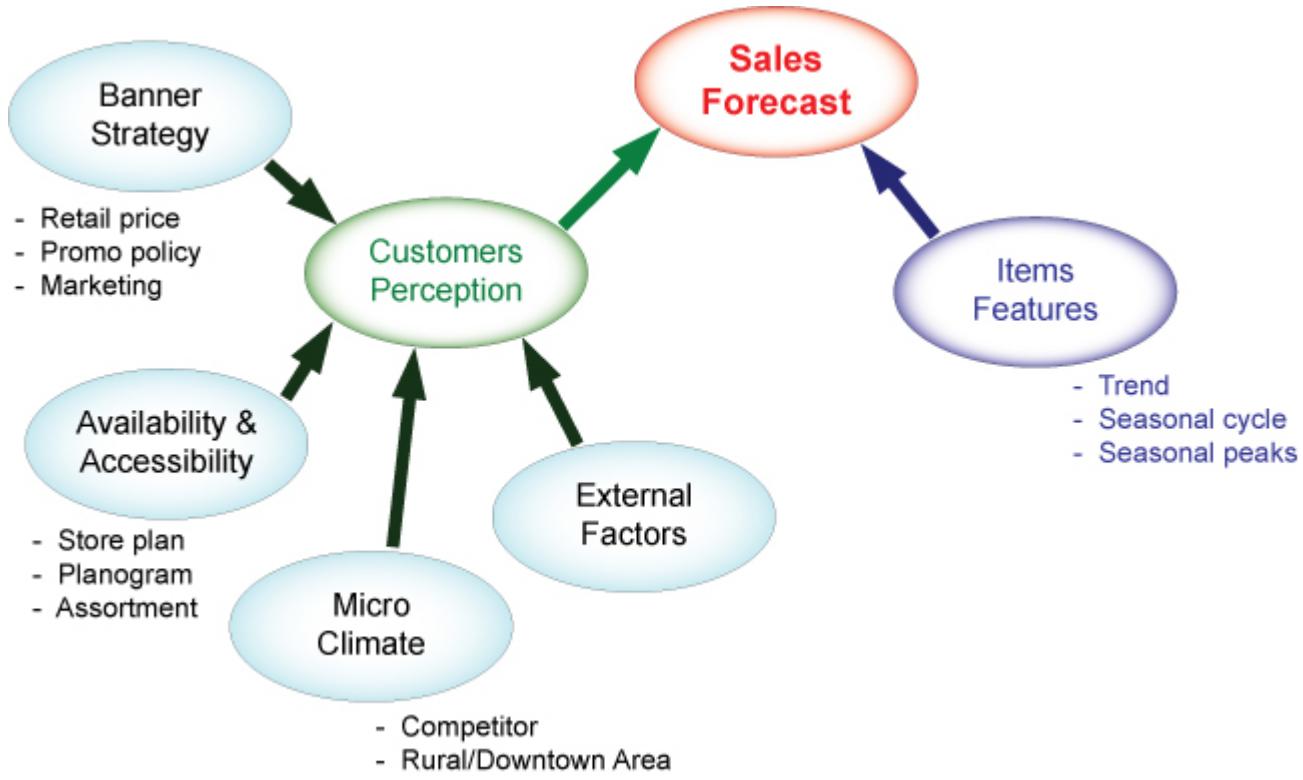


Fig. 2. *How to explain a sale?*

Sales quantity is an easily and available data from which information can be extracted to explain the sales. This approach, called modeling, is the one used in *Ibs* - Analytics for forecasting both regular sales and sales under special strategic rules.

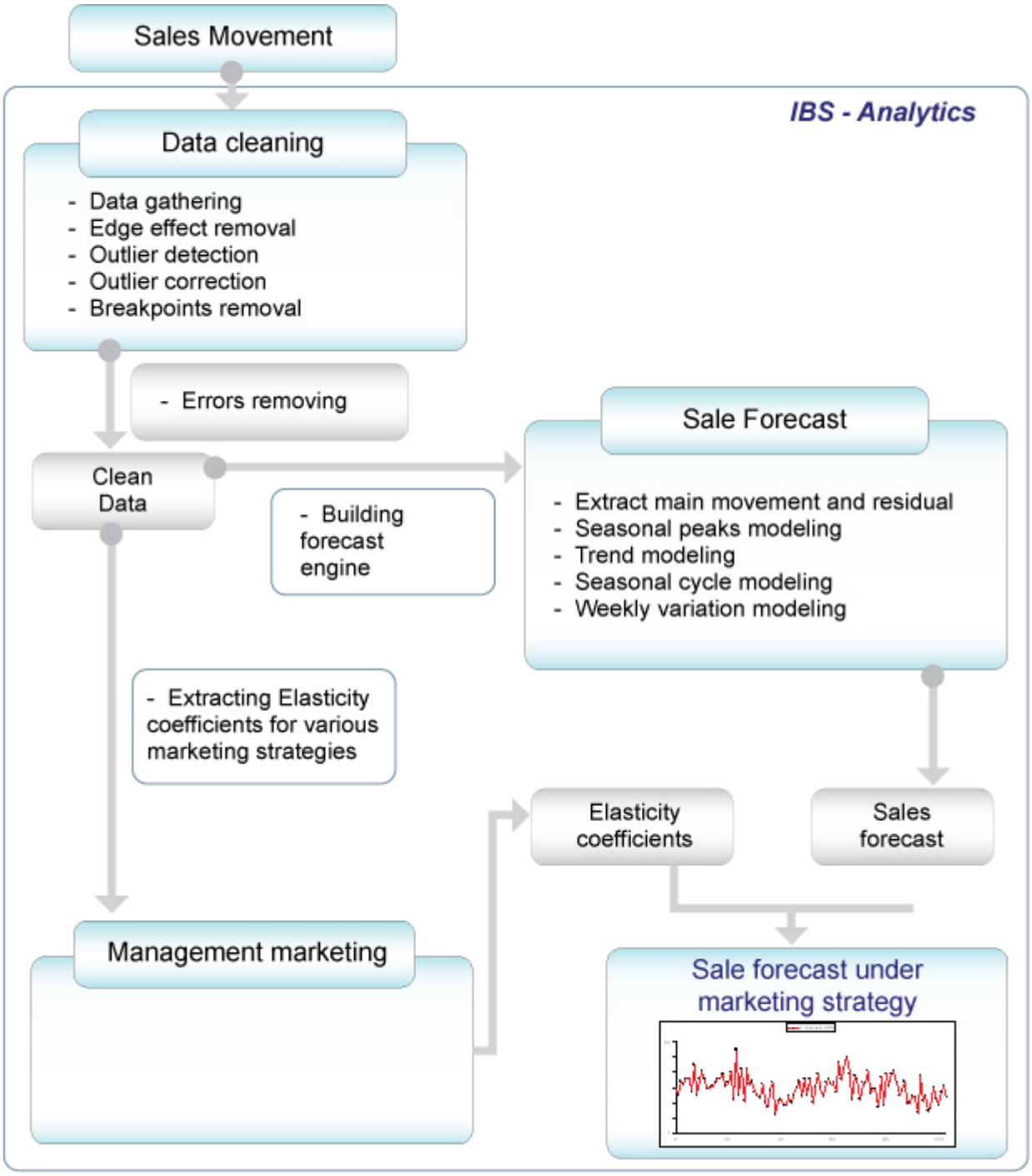
The goal of modeling in retailing is to be able to explain the quantity of item sold by the most significant variables aforementioned in order to build a model able to forecast the sales in various scenarios in the future. *Ibs* - Analytics uses the history of sales data to extract significant information such as the seasonal cycle, the seasonal peaks and the trend of the sales in order to provide a baseline of predicted sales.

2.3. Overview of the general modeling process

Ibs - Analytics modeling process is divided in three (Fig 4):

- Data cleaning which aims to increase the data quality by removing erroneous events (next section)
- Sales forecast engine, which will predict the regular sales (available in a dedicated white paper. Visit our website **www.ibs-softsolutions.com**)
- Business Marketing Management, which applies the effect of various marketing strategies on the sales available in a dedicated white paper. Visit our website **www.ibs-softsolutions.com**)

Fig. 4. *General overview of Ibs - Analytics process*



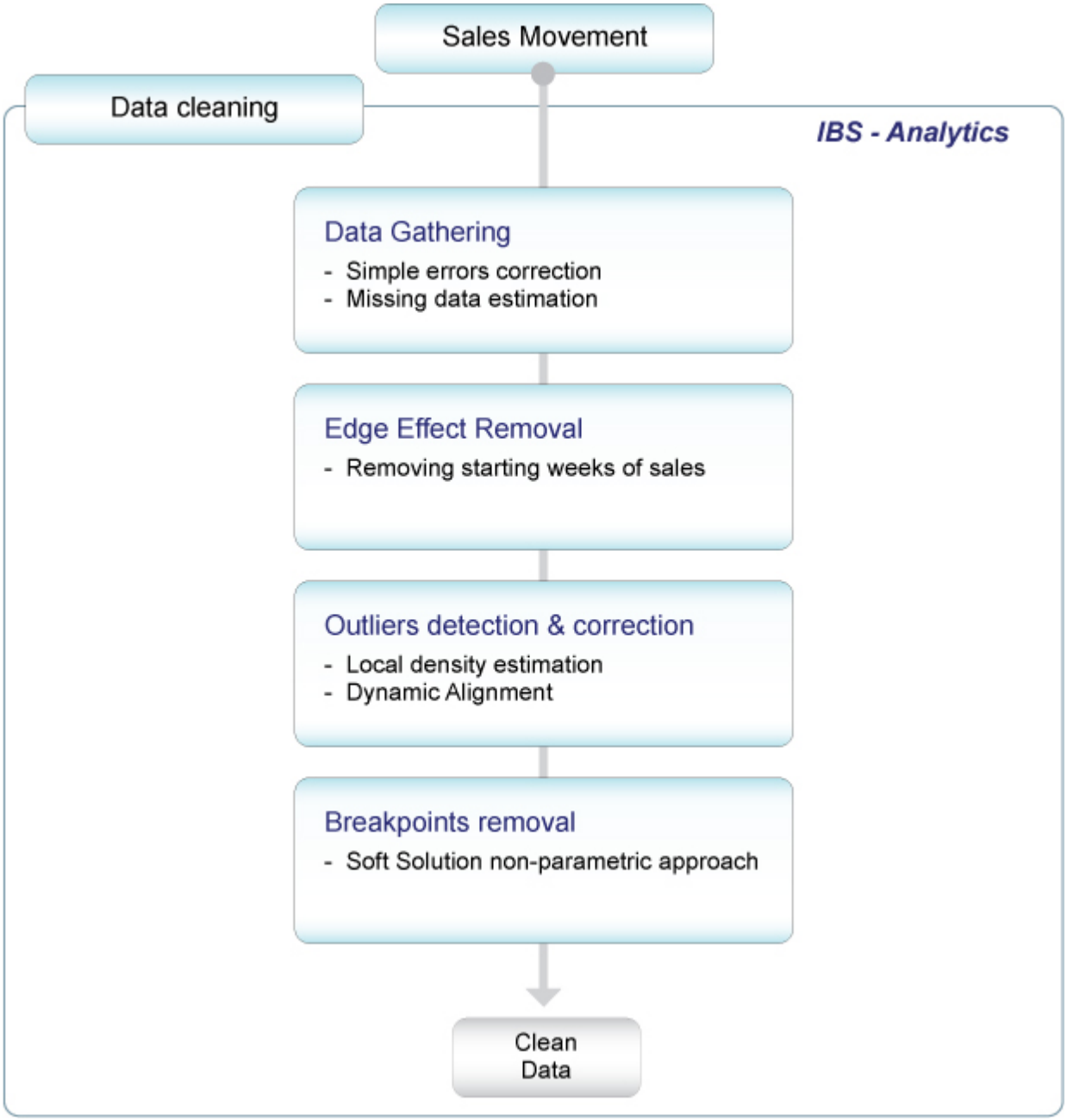
3. Data cleaning

The whole process starts with raw data coming from one or several databases, which consists in the sales history of items by weeks, by stores and other information. The goal of this step is to obtain for each items cleaned data for the next learning parts; this is a fundamental phase as data mining from erroneous data leads to less accurate model. Cleaning data implies to remove and correct anything that deviate from a stable state.

3.1. Data cleaning workflow

Global overview of the data cleaning process is shown Fig 5:

Fig. 5. *Data cleaning workflow*

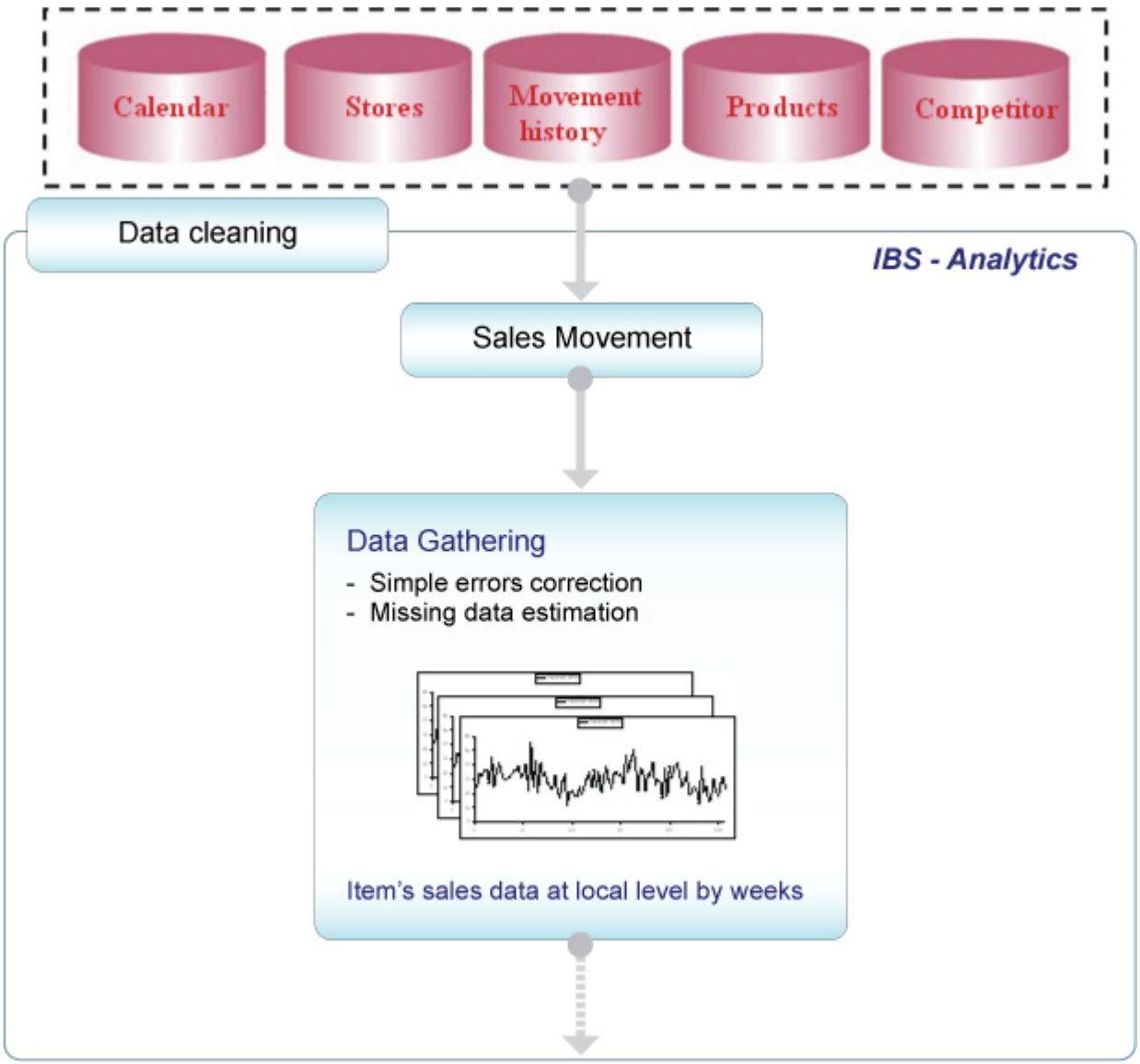


3.2. Data gathering

This first service performs initial data gathering of the data (Fig 6), with basic rules for simple errors correction like change for mistyping (i.e.: ‘o’ and ‘l’ are respectively changed in ‘0’ and ‘1’) and missing data estimation based on logical inference and local estimation methods. Then, for each item, sales data are weekly added up from the main movement to a local level (store, region).

Working on such specific area size allows us to work on bigger and more stable quantity; it also avoids modelling small quantity, which can be very sensitive.

Fig. 6. *Data gathering*



3.3. Edge effect removal

Using sale data history, it may happen that some products start being marketed in the middle of the global history. Such events, called Edge Effects, have to be removed in the second part of *Ibs* Analytics process before modeling as only the stable part of the sale history of an item is representative (Fig 7).

Ibs Analytics' approach is to remove the first six weeks of sale data when a new item is being marketed.

Indeed, the first and last weeks of item marketing are not relevant in predicting future regular sales. Thus we obtain more stable and reliable data for the remaining process of *Ibs* - Analytics.

Fig. 7. *Non-pertinence of edged effect*

Local density estimation:

Local Density Estimation is a method, which computes the distances between every points of the main sale event.
Then for each point, local density is estimated.
Finally, for a point, if the ratio of his density over the mean density of his neighbours is under a defined threshold, this point is spotted as outlier.

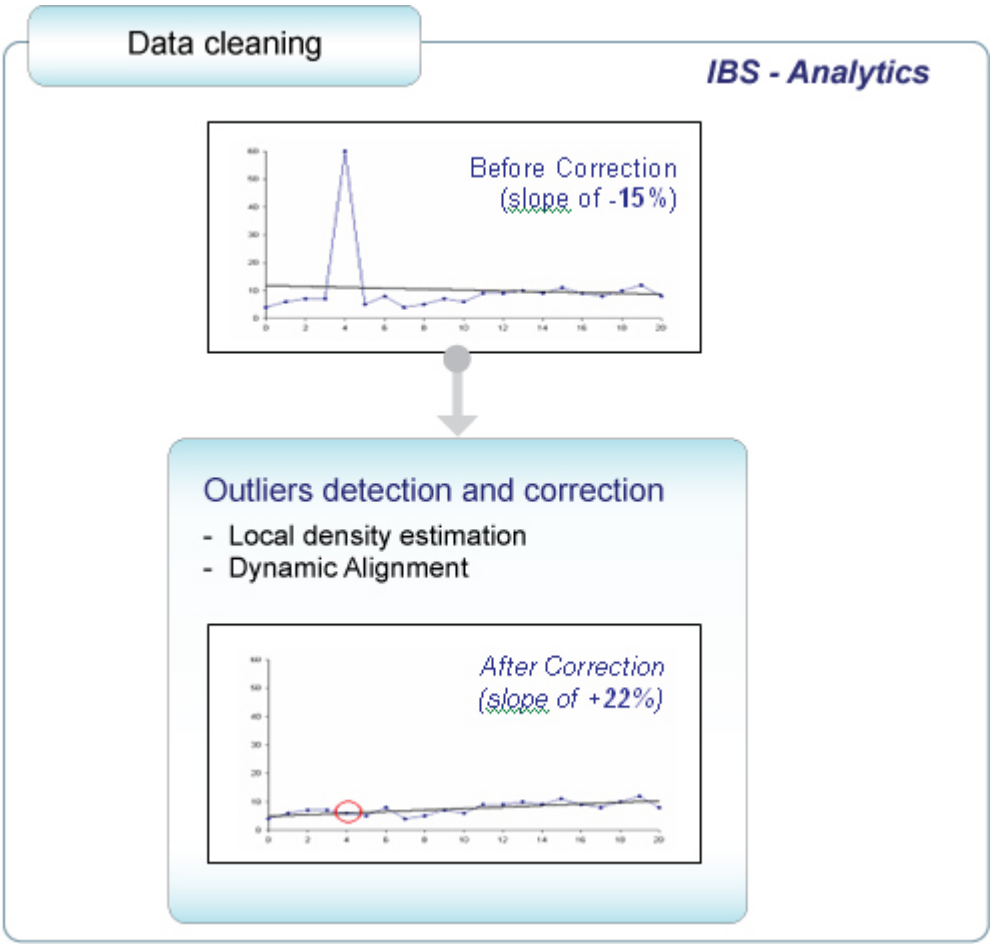
3.4 Outliers detection and correction

An outlier is a value that significantly differs from the others. In data mining, one does not want to use these data, as there is no explanation for such difference in the model and thus it will disturb the learning phase of modeling, which will lead to reduced accuracy.

Ibs Analytics implements a spatial approach to detect outliers called Local Density Estimation. This method has been fit from scientific research (Breuning & al. 2000) to the retails sales specifics. Facing classic statistics, this approach is more dynamic and produced better result to detect outliers as its local approach better adapts to change in sales.

Outliers' detection is also useful to detect seasonal peaks, which are defined as reoccurring events at a given frequency. *Ibs* Analytics store the weeks when the events happen, as it will be used to adjust seasonal peaks events on regular sales forecast based on the Dynamic Alignment methods (Smith & Waterman 1981).
By removing this data and replacing it by logical inference and local estimation (Fig 8),
Ibs Analytics avoids inserting bias in the forecast engine computation

Fig. 8. *Bias introduced by outliers*



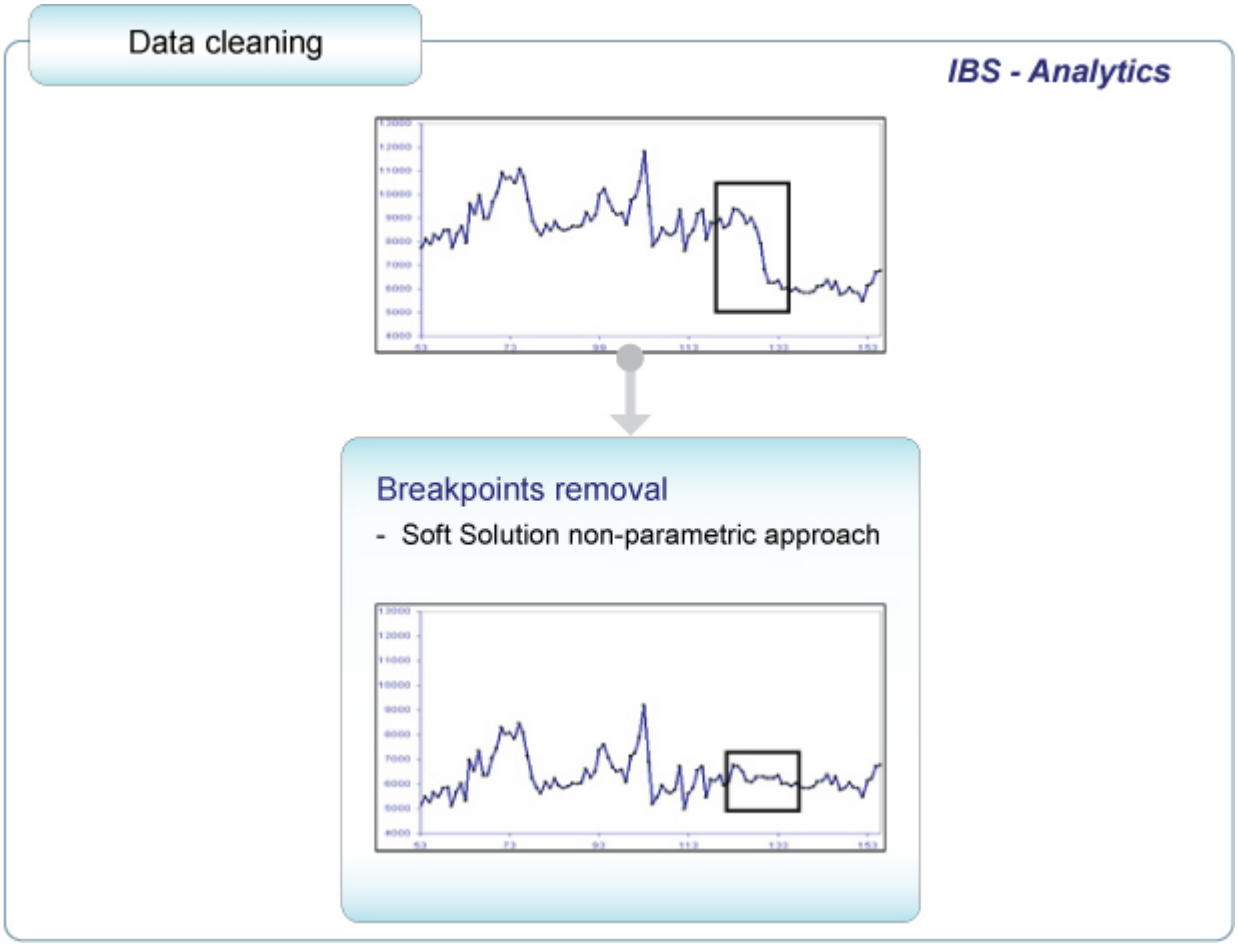
3.4. Breakpoint removal

A breakpoint is an event when sales abruptly change, in a short amount of weeks, from a stable sale condition to a superior or inferior one, which is also stable. This behavior can be explained in many ways; for example, the item is now sold in twice the number of stores or space planning allows more room for this item. Anyways, as there is not always data to explain these events, they become not representative for regular sales forecast.

Soft Solutions has developed for *Ibs Analytics* a new non-parametric approach for Breakpoint detection and Breakpoint correction. For each item, data history is truncated into stable part. When two stable parts are separated by a short decrease or increase, a break point is highlighted.

Finally, like in the previous parts, learning from such data will lead to erroneous results where similar pattern will be displayed. Thus *Ibs* - Analytics corrects these events by adjusting the oldest part at the sales level of the newest one (Fig 9).

Fig. 9. *Breakpoints correction*



4. Conclusion

Soft Solutions has developed a dedicated module *ibs* Analytics for information mining and retail sales forecast to offer a decision-making tool by anticipating consumer's need. It implements concepts and robust methods from scientific researches and tuned them for the retail business.

As described in this specific white paper, *ibs* Analytics integrates a complete and advanced cleaning engine. This one is the first step of the modeling process and is a key-point in order to achieve accurate modeling.

Combining techniques from statistics and data mining (local density estimation), ibs Analytics is able to detect any kind of deviant data within sales histories and adjust level of sales to ensure providing consistent data to the forecast engine.

References:

- M. Breunig, H-P Kriegel and RT. Ng and J Sander, 2000, "LOF: Identifying Density-Based Local Outliers"
- H. Krim, D. Tucker, S. Mallat and D. Donoho, 1999, "On Denoising and Best Signal Representation"
- H. Jair Escalante, 2005, "A Comparison of Outlier Detection Algorithms for Machine Learning"
- S. Walfish, 2007, "A Review of Statistical Outlier Methods"

For more information:

- **Website:**

  - Soft Solutions
    http://www.ibs-softsolutions.com
  - Soft Solutions – Analytics
    http://www.ibs-softsolutions.com/eng/Retail-Analytics-Suite.html