Written press logistics

How clustering and sales forecastingcan levitate an industry's distribution

Asmaa Sabiri[1], Fouad Riane[1,2]and Alami Semma[1]

[1]Mechanical Engineering Research Laboratory, Facultyof Science and Technology
Settat, Morocco
[2]Ecole Centrale Casablanca
Casablanca, Morocco

***Abstract* – Written press has been, and still is in Morocco, an important means to communicate news and knowledge.Press releases still have their advantages and devoted fanatics. Nevertheless, a general decline may be noticed worldwide due to digital substitution and general reading decrease. Moroccan press undergoes additional hard time as its management systems struggle to cope with a fast-paced industrial development and network modernization.**

This paper is an inception study. It tackles reflections and insights about Moroccan pressdistribution systems and suggests tailored solutions accordingly. Based mainly on sales analysis,we will be discussing sales points clustering and sales forecastmodeling. Hopefully,it will help understand thisindustry's complications and puzzle out distribution and marketing possibilities.

Keywords – distribution; forecasting; newspapers; seasonality; sales; hierarchical clustering; returns

I. INTRODUCTION

The lifecycle of a newspaper begins with several ideas, written virtually on software then printed physically on a set of sheets. Once it is created, the newspaper begins its journey to perish; it will be seized from the printer, transported to the press mailing society, delivered to vendors then read by customers. The problem is that daily press has very tight time constraints. If the newspaper is not on the sales point shelf early in the morning, the opportunity that it can be read is missed and the newspaper can be considered obsolete.

Editors are responsible for producingthe content, printing itand conveying the newspapers to the mailing society. The latter is in charge of distributing the newspapers to a large network of sale points, all over the country. Ranging from groceries and street vendors to newspapers shops and libraries, all have to be provided daily with a different amount from each newspaper.They can be located in urban areas, rural regions, near locations or distant and hardly known places. Depending on locations, and networks' nature, newspapers can be distributed differently, using different transportation means and delivery accommodations.

The amount of newspapers to be delivered to each sale pointis defined by the mailing society on a daily basis. At times, it can also rely on the editors' recommendations out from newspapers' content and location-based reading expectancy.

In addition to the main flow, resulting in the newspapers being delivered to sale points, a return flow is also managed by the mailing society. A sold newspaper is out of their business, but an unsold one is not. It has to be transported from vendors as a return back to the depot, stored then retrieved later by editors or destroyed. No doubt this double flowprocess should be optimized, each flow with a different logic, as it is daily repeated and any flaw within its operations will snowball indefinitely through time.

When a newspaper is being operated, costs include manpower, transport, facilities, data storage and many indirect costs making it crucial to reduce the returns and therefore optimize the number of newspapers distributed. This seems legitimate as long as sales are not affected, which may miss the very purpose of the activity.

Actually, editors always seem to want a high amount of copies to be distributed. They suggest that theamount distributed to sale points is heavily correlated with the amount of sales. It complies with the fact that missed sales decrease with overestimated distributed copies but returns will be more costing eventually.

Therefore, the first conflict appears; editors wanting to maximize coverage for their titles and mailing society wanting to optimize the number of distributed copies. Finding the right balance may be possible by confronting costs and sales revenue, when adding a certain amount of copies, but that balance may be inconstant, hardly accurate and therefore expensive to assess.

A second conflict that may be confronted is that the mailing society has also to manage special political or governmental recommendations; regarding press presence or newspapers coverage. It is part of its duty to bow to any action related to raising population awareness or spreading important information.

Overall, it is no secret that the press industry decline is all but reversible. Therefore, it is more than urgent to master technical and operational issues in order to have a clearer vision on how to encounter the most imminent and overwhelming ones.[1]

The generation of distribution orders aims to define the number of copies of a given newspaper that must be routed to each sale point. Sales forecasting is a key element for this purpose. It is based on the sales history along with other variables (weather, special events, sporting events, political news, etc.). Come also into consideration some rather constant

Xème Conférence Internationale : Conception et Production Intégrées,CPI 2015, 2-4 Décembre 2015, Tanger - Maroc.

Xth International Conference on Integrated Design and Production, CPI 2015, December 2-4, 2015, Tangier - Morocco.

data, such as consumers' behavior near each sale point and the characteristics of the distributed products.

The aim of our research is to make the generation of distribution orders more automatic and reliable.We will proceed initially to a sale points' categorization. This will allow us a better understandingof their behavior in order to provide a tailored calculation method of the delivery amount (i.e. the number of copies of a given newspaper to be delivered to a sale point). The suggested method should be easy to understand by the user anddepending on minimalsetting parameters.

II. Clustering

Clustering is one major issue to study ahead of making an effective forecasting model.

Forecasting is about estimating the right amount of copies to a set of salepoints. However, for the titles and the points of sale to be the right ones too, a proper categorization is needed. Newspapers have different characteristics that can be considered for this aim; such as price, language andcontent-based targeted readers(Women, children, sport fanatics, specialized readers, etc.).

On the other hand, the categorization of sale points is much trickier. Sale points' characteristics are more divergent and hard to define objectively, such as localization, foot traffic, opening time, main activity, customers' habits and traits and shelving potential.

Based on the previous characteristics, a fine model may be build up in order to assign new or existing titles to points of sale, to forecast sales more accurately and to orient marketing actions.There are different possibilities of how sale points may be categorized, we chose to perform a statistical method in this paper.

A. Activity-based Qualitative Clustering

As a start, along with experienced professionals, we considered sale points' main activity and made an qualitative clustering based on the expected sales' general trend. It resulted in six categories:

- Sale points or individual 'diffusers' serving exclusively subscribers. This category needsneither shelves nor a forecasting system. Their customers may be individuals or companies. Their sales are constant at any time given of the year.
- Newspaper shops or libraries, serving essentially subscribers (more than 80%). Their sales are almost constant. They may not serve many regular customers but are more likely to do so given their activity.
- Day street vendors. They have their own distribution network. Their sales are highly influenced by weather changes, special events and delivery lateness. They may be perceived as less organized than shops, yet they provide a superior revenue. Their sales predictability is the most difficult though. Their customers may vary between regular ones or non-declared subscribers.
- Night street vendors. They are similar to the previous category except that they have a special delivery schedule.
- Sale points with any given main activity, disposing of shelves, and serving regular customers.

In order to make delivery forecasting, we will be using mainly the three last categories, being the least predictable and therefore, more critical. We assume that customers all over the country have similar behavior and proceed to test a forecasting algorithm accordingly.

B. Hierarchical Clustering

Hierarchical clustering is one commonly used method. It providesdecent visuals and allows easy-to-read clusters. On the other hand, it has a bit of a complexity; before exploiting the results, one has to be aware ofhow to organize data before clustering, to choose the appropriate distancemethod and even howto customize plot colors.[2]

In order to make sale points clustering, we considered a year of sales records, registered by title, day and sale point. We will be using hierarchical clustering in R-Studio software (3.2.1 version) on windows 8.1.

At first, we cleaned data from subscribers and special sale points. Then we tested a first clustering based on newspapers titles, the results were hard to read and rather confusing when it comes to identifying each cluster apart. For the latter reason, weaggregated sales by content classes; resulting in seven of them: economy, politics, sports, family, culture, children and other.

Using R console, we created first a data table, then converted it to a distance matrix by the Manhattan method [3] creating distances as the sum of absolute differences. We applied hierarchical clustering using Ward method, considering the error sum of squares [4] for clustering. The resultsare shown in Fig. 1.

The hierarchical clustering gives a tree diagram with a countable number of cluster possibilities. The number of clusters can be determined by statistical methods that have not been discussed in this paper. We manually tested several potential possibilities instead. While fixing the limit of clusters agglomeration (Red rectangles) at 300 copies sold, we can deduce six main clusters of sale points. This gave us a reasonable number of clusters with the least clusters volume variability.By choosing a higher limit, the number of clusters will decrease and thus produce bigger clusters that may be confusing. By choosing a lower limit,the number of clusters will increase, creating even more smallheterogeneous clusters.

The clusters are rather recognizable but we cannot tell what content classes are included in each cluster and at what amount. This is where a heat-map will come in handy.

Using the same software plotting feature, we represented the allotment of content classes on clusters. The result is shown in Fig.2.

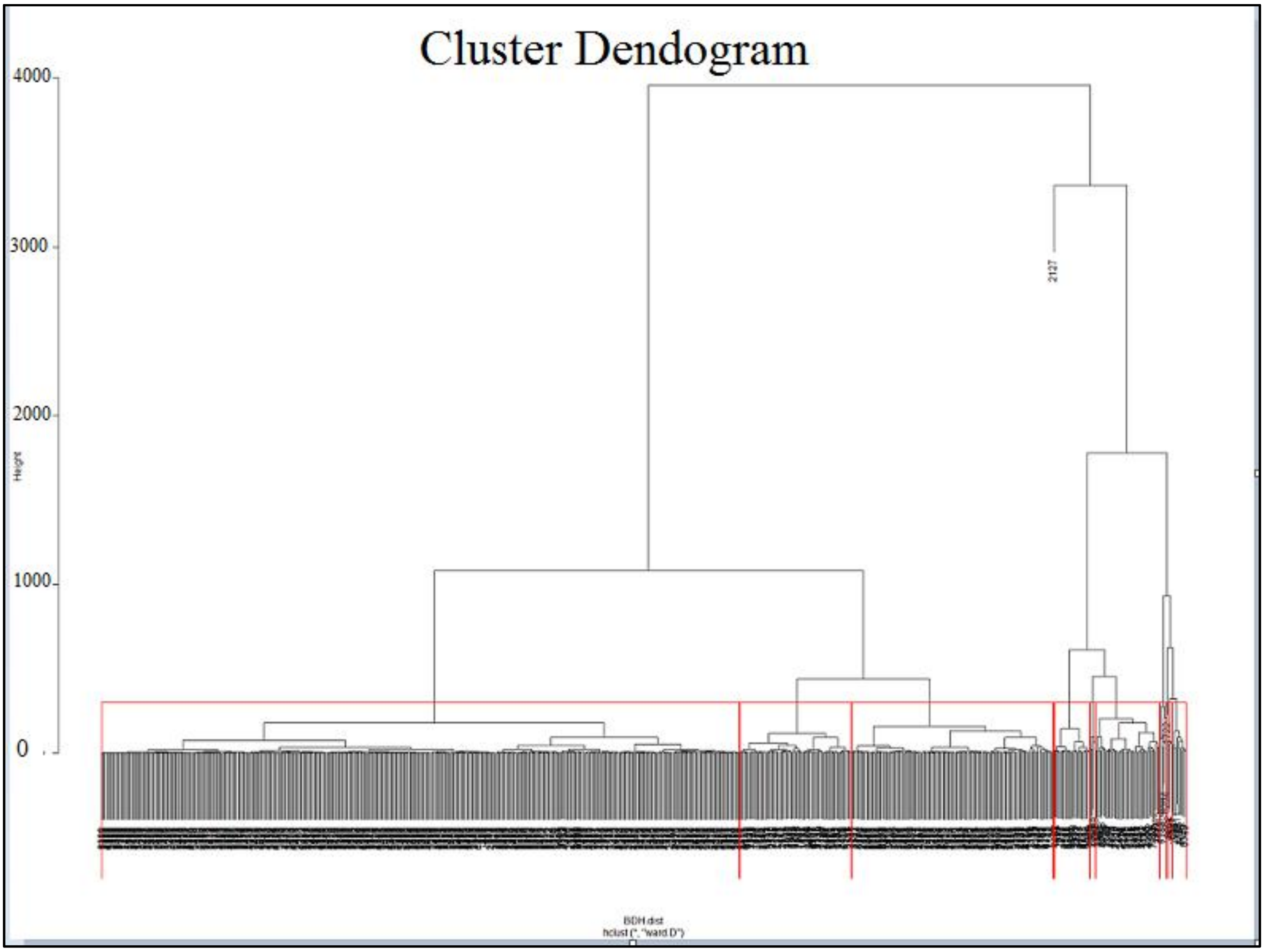

Fig. 1. Dendogram – Hierarchical Clustring

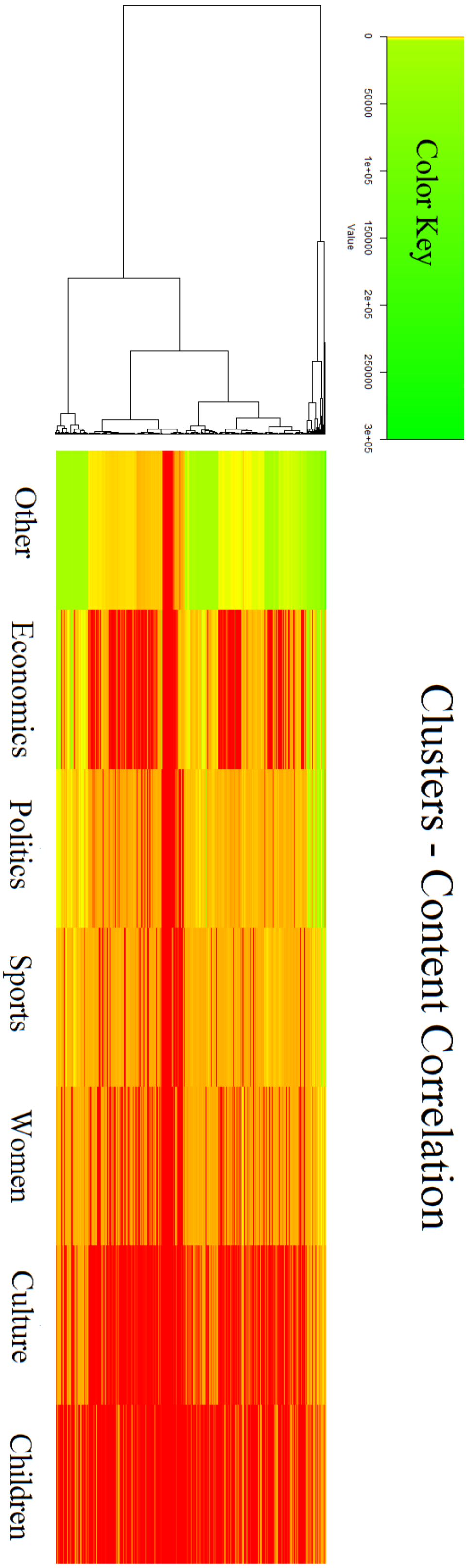

Fig. 2. Dendogram – Hierarchical Clustring (Heat-Map Correlation)

As for this representation, we can easily identify correlation between content classes and clusters. The aim of this categorization is to deduce content clusters composition then associate sale points with the resulting content clusters.

Associating sale points to content willmainly help associate new titles with existing sale points, based on their related cluster. Optimizing this process will help logistics addressing the right newspaper to the appropriate point of sale and thus surely reducing returns related to bad product positioning.

While knowing which newspaper to deliver to which point of sale, we still need to determine the right amount of copies to ship to the latter.

III. Delivery Forecasting

In this paragraph, we will be discussing distribution orders based on sales/delivery forecasting. A distribution order is basically the number of copies from each product that has to be delivered to each sale point. The sale points' distribution orders are aggregated by regions and cities in order to determine the volume to be routed to each location and the transportation mean that will deliver the products.The mailing society is responsible for this calculation.

Distribution orders are based mainly on forecasting sales. They are generated for each sale point based on its sales forecasting and then adjusted according to the product characteristics.Two forecasting models are discussed below, resulting each in a different outlet.

A. Importance

Sales forecasting is one major operation that the newspapers' industry relies on. Assuming all strategic decisions are made, it all comes down to the process of assigning one number, related to one title, to one sale point. This process depends on a set of information that has to be accurate, up to date and in the right level of details.[5]

Missed and overestimated sales are two important indicators for sales forecasting and generating distribution orders. The first one results in lost sales and the second in high returns.

The needed information for distribution orders comes from sales records, special events, weather, newspapers and points of sale characteristics, time of the year and list goes on. The most important dataset that we will be relying on, while building the forecasting algorithm, is the sales records because it is inconstant and says a lot about the efficacy of the forecasting system.

It is also important to know that one number from the sales record dataset does not always have the same meaning nor the same weight. Assuming that the number "1" representsone sold copy in a relatively well-positioned area, with satisfactory sales statistics, related to a title and to a point of sale. That exact number is far from being the same as a "1" sold copy related to a point of sale that nearly sales onecopy a day of the same title. Carefulness is to be considered while handling even the simplest data.

B. Case 1 : Returns Optimization

The main goal, as it has been mentioned previously, is to find the right amount of newspapers to deliver to each sale point. We will refer to it as the "Delivery amount". The latter is to be necessarily deduced from the forecasted sales.

In our case study, defining the delivery amount is similar to facing a melting pot of different settings and variables of which no one holds the exact tuning, able to produce the right results. As a start, we fixed at random different parameters in order to simplify our study and try to build a raw but scientific procedure. The mailing society currently considers a simple average for forecasting associated with some settings related to each product.

We will be assuming that the sales are predictable and make the simplest calculation process possible in order to forecast sales andthen let the resultsguide us through any customization of the original formula.

We considered at first optimizing the daily delivery amount ($D_{d,w}$) resulting in an algorithm that helps calculating first sales forecast ($SF_{d,w}$), at a given day (d) in a given week (w) then add an amount (A) to deduce the delivery amount ($D_{d,w}$), as formulated in (1).

$$D_{d,w}= SF_{d,w}+A \qquad (1)$$

We tested pseudo out-of-sample sales forecast with several common time-series analysis models (linear regression, moving average, exponential smoothing, median smoothing seasonal composition, etc.). Then we compared their results based on two main indicators: returns proportion and missed sales. Although they had very similar results, smoothing historical data based on median operator seemed to be the best choice because its implementation is simple and it does not take to consideration aberrant values that have to be overlooked anyway. Hence, the retained sales forecast formula will be the median of sales ($S_{d,x}$) observed in every similar day (d) from the previous n weeks (x ranging from 1 to n), as in (2).

$$SF_{d,w}=\text{Median}S_{d,w-x} \qquad (2)$$

The number of weeks to be considered depends on the deviation of the seasonality. The larger the deviation, the longer the span.

The amount (A) isformulated in (3). It has been established in such a way as to deduce the delivery amount from the sales forecast autonomously and without any previous settings, regarding the characteristics of the sale points or the newspapers.

$$A = \left(1-\alpha_{d,w-1}\right) \times k \times SD\ (S_{d,w-x}) \qquad (3)$$

(A) stands for the amount to be added to the sales forecast.$\alpha_{d,w-1}$ is the proportion of returns of a similar day from the previous week.SD($S_{d,w-x}$)is the standard deviation of the sales of similar days from the previous n weeks (x ranging from 1 to n). Finally,k is set to calibrate (1), adjusting the deviation in (3)with sales in (2) (ranging from 1 to 3).

The expression [k× SD($S_{d,w-x}$)] adds a reasonable amount to the sales forecast based on the sales variability. It will be added entirely if there are no returns ($\alpha_{d,w-1}$=0) which probably means missed sales, or partially in proportion with the returns, or not at all if the there are no sales registered ($\alpha_{d,w-1}$=1). An example of sales forecast based on the described model is illustrated in Fig. 3.

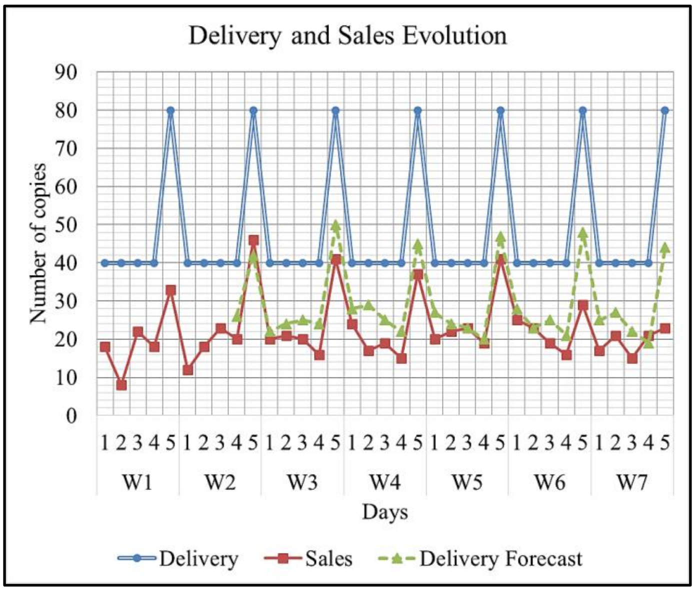

Fig. 3. Evolution of delivery forecast compared to actual sales and planned delivery (Algorithm 1).

We can easily take notice of how the forecasted delivery curve is maintaining a minimum gap with sales. As for the point "day 4 in week W7", where sales seem to be superior to the delivery forecast, sales would have been missed that day for a reason, which is difficult to avoid taking into account the variability of sales and special events that may affect them and that are not being considered in the established algorithm.

Let us take note however that sales returns represent 17% of delivery and are reduced by 35% compared to the realized/planned delivery actually doneby the mailing society at the period. This decrease represents a fair amount given the range of the sales numbers (13 to 48 sales).

No doubt, this algorithm has certain limitations. When wanting to increase the delivery amount, the result of (3) tends to over-fluctuate the results. Hence, it cannot be intended but to optimize the returns. We assume that there is a seasonality. It is indeed present for the majority of clients but may not be for others. As for miniscule or rare sales, it is also difficult to get the right delivery amount. Delivery in this context should be fixed differently. This issue will be discussed later in this paper.

Overall, this algorithm allows a dynamical optimization of the delivery amount based on the sales variability and calibrated by returns. It is not always the case though; while optimizing sales may avoid returns, missed sales are more likely to happen. As for the case where returns are less costing than missed sales, returns (desired proportion of it) are to be fixed and the delivery amount is to be set to reach it. This case is more likely to occur because editors worry less about returns. The following algorithm describes sales forecasting with a fixed proportion of returns.

C. *Fixed Returns*

Let us bring back the fact that sales forecasting is a process executed on a daily basis. Therefore, a main challenge is to make the algorithm as smooth and simple as possible in order to be performed rapidly. An additional constraintoften occurs, requiring swiftness; editors may deliver a different number from the one declared to the mailing society. Therefore, delivery amounts for each client have to be changed. This process ought to be as quick as possible. Otherwise, delivery will be deferred and sales will be missed.

The proposed algorithm calculates the delivery amount deduced from sales forecasting. The latter is calculated based on a moving average adjusted by seasonality.

The calculation of the moving average $MA_{d,w}$, for a given day (d) in a week (w), is based on thesales ($S_{i,w-t}$) of n weeks before the forecast, as in (4). The term (i) indicates a given day of the week and (m) the number of delivery days per week.(w) indicates the actual week of the forecast and (n) the number of averaged weeks browsed by (t).

$$MA_{d,w} = MA(S_{i,w-t}) \begin{cases} 1 \leq i \leq m \\ 1 \leq t \leq n \end{cases} \qquad (4)$$

The sales'seasonality coefficient$Se_{d,w}$, for a given day (d) in a week (w), iscalculated following a three steps process:

*a) Raw seasonality coefficient $Se''_{d,w}$:*For a given day (d) of the week (w), raw seasonality coefficient is computed based on sales ($S_{d,w}$) of the same day and on the related moving average ($MA_{d,w}$). The formulation is described in (5).

$$Se''_{d,w} = S_{d,w} \div MA_{d,w} \qquad (5)$$

b) Corrected seasonality coefficient$Se'_{d,w}$: will be calculated, as in (6), based on raw seasonality coefficient $Se''_{d,w}$ in the previous (m) delivery days per week.

$$Se'_{d,w} = \frac{Se''_{d,w}}{\sum_{1 \leq i \leq m} Se''_{i,w}} \times m \qquad (6)$$

c) Seasonality coefficient: The estimated seasonality coefficient ($Se_{d,w}$) for the day we want to forecast will be computed as an average of the corrected seasonality ($Se'_{d,t}$) of the (p) previous similar days :

$$Se_{d,w} = \frac{1}{p} \times \sum_{1 \leq t \leq p} Se'_{d,t} \qquad (7)$$

The sales forecasting ($SF_{d,w}$) will be then the multiplication of the moving average ($MA_{d,w}$) by the seasonality ($Se_{d,w}$), as in (8):

$$SF_{d,w} = MA_{d,w} \times Se_{d,w} \qquad (8)$$

At last, the delivery amount ($D_{d,w}$) will be computed as in (9), dividing the sales forecasting ($SF_{d,w}$) by $(1-\alpha)$, α being a fixed parameter computed based on the proportion of returns realized the previous days.

$$D_{d,w}=\frac{SF_{d,w}}{(1-\alpha)} \qquad (9)$$

Fig. 4 and Fig. 5show two examples of the previous algorithm; targeted returns are fixed at 35%.

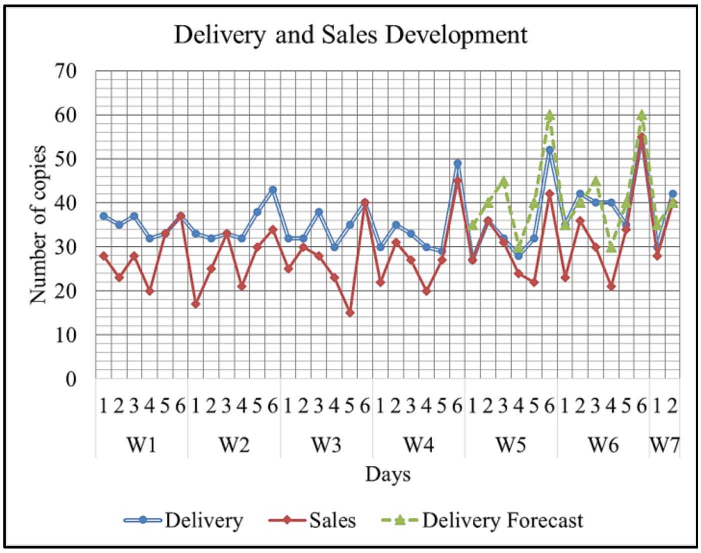

Fig. 4. Evolution of delivery forecast compared to actual sales and planned delivery (Algorithm 2 – Example 1).

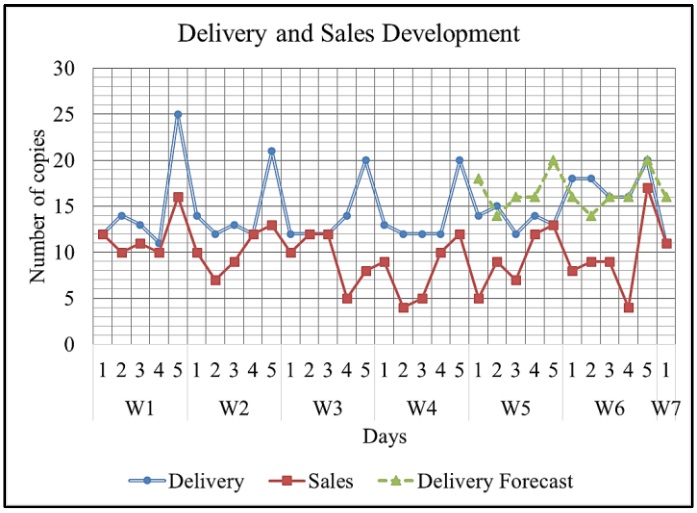

Fig. 5. Evolution of delivery forecast compared to actual sales and planned delivery (Algorithm 2 – Example 2).

In both examples, forecasted delivery inducesless but more stabilized gap with sales,based on the targeted amount of returns.

For the second example, realized delivery produced some missed sales (W1,1 ; W2,4 ; W3,2 ; W5,5 ; W7,1). The forecasted delivery would have been able, if realized, to avoid missed sales and reduce returns on 5th days.

The delivery amount is rounded by multiple of 2, 3 or 5 relatively to a low, medium or high average of sales (2 in the previous examples). Besides the fact that it will accelerate the counting process of newspapers at the departure, there is no need to over fluctuate the delivery amount while most of the calculations trims from integer numbers.

The limitations of this algorithm are mostly the same as the previous one.

We tested the described algorithm on 30 sale points from the categories detailed in II.A., for three different products over one week of distribution. The population size has been limited mostly by technical complexities.

The results are summarized in TABLE I.

TABLE I. ALGORITHM 2 – TEST RESULTS

Product	**Test results**			
	Returns proportion		***Missed sales rate***	
	Conventional method	***Algorithm 2***	***Conventional method***	***Algorithm 2***
A	46%	39%	17%	9%
B	25%	24%	22%	15%
C	46%	38%	22%	12%
Summary	39%	34%	20%	12%

We can clearly identify that returns proportion has been stabilized, decreasing over-estimated ones in order to reduce returns (Products A and C) and decreasing underestimated ones in order to avoid missed sales (Product B).

While analyzing tests, we could discern a potential categorization based on the average sales. It could help determine the targeted returns amount by analyzing clients with close sales' levels similarly as it has been perceived that they have similar behaviors [returns and missed sales] regardless of their activity or the content they sale. It is yet to be thoroughly studied.

D. *Rare and sparse sales*

Sale points with rare or very few sales are sure to be processed differently. Firstly, because there is nearly no apparent seasonality. Second, because they seem to have utterly unpredicted sales, for they have no identifiable pattern.

At the moment, three categories have been deduced. TABLE IIbelow summarizes them:

TABLE II. SPARSE SALES CATEGORIES

Categories	**Sales Evolution Characteristics**		
	Variability	***Maximum***	***Incidence***
A	High	Irregular	Low
B	High	Moderately consistant	Medium
C	Low	Consistent	High

Following these non-exhaustive categories, delivery computation will vary according to each case scenario. It may be a fixed amount if the maximum is consistent and the variability is low. It may also be calculated based on the probability of incidence if the latter is low.

Calculations resulting in small integersare not the best conditions though for advanced calculations. Important information may be lost while rounding numbers. The first step would be to consider an equation that helps identify the different categories in order to relate them to the matching scenario. The rest should be easier as tests are surely made among the right population, enabling a clearer results analysis.

IV. Conclusion and prospects

The established algorithms have been quiteaccurate. The most relevant issue is how to tailor them according to the activity-based clusters.It still needs polishing regarding clustering analysis association, in order to identify tailored settings for each cluster.

As for the accuracy, sale points that have relatively high loyal customers, or serving subscribers, need to be allowed less returns because their sales trend is more stable.On the other hand, street vendors are more affected than the rest by weather conditions, specific events and other circumstances. These events must be studied thoroughly in order to be modeled and automated. There is a need of many years of consistent sales' records in order to identify the related trends and seasonality, which isnotthe caseunfortunately.

Additional constraints related to the environment of the sale points are also to be considered; being in anindustrial neighborhood, the sales drop at the weekend, in opposition to residential neighborhoods where sales usually rise during the same time.

Forecasting for spare sales is at its debut and appears to yield more constraints and complexities. But its resolution depends mostly on how the society is willing to bet on these sales points and how much return's rate it can afford. Editors may also have to decide whether they are interested mostly in sales or just in theshelving.

Clustering methods are countless and each one may give different clusters on every execution, based on initialization and other methods' settings. They have to be repeatedly tested and thoroughly compared. It is also important to consider that clustering sale points may be intended to help organizing marketing actions, in order to makeclusters accordingly. [6]

For the time being, a study is conductedin order to assign products to sale points. For the clusters to be more meaningful, we have to make sure that every sale point is assigned every product that itmight possibly sell. For each sale point, products that have never been assigned, or have been assigned but did not work well afterwards, have to be identified.

A benchmark is more than welcome and a thorough knowledge of products content will surely help but is not sufficient to settle; as products with the same content may have different success rates. This study will supposedly give combinations of new items for the sale point to choose from. After being assigned, products' sales must be then observed closely and the assignments confirmed.

Having available data has been a huge push up to our study. An important issue has to be considered though. As we are advancing, some data seemed to be misleading, as it does not represent what we assume it does e.g., when a sale point is closed and has a return proportion of 100%. Not considering that returns are in this case justifiable, it may induce false results.[7]

Therefore, associating and verifying results based on field observation, rather than just numbers is crucial. It surely reveals a lot about numbers signification.

Acknowledgment

We thank the Moroccan mailing society that provided us with all the data we needed to accomplish this research.

References

[1] Rick Edmonds, Emily Guskin, Amy Mitchell and Mark Jurkowitz. "Newspapers : Stabilizing but still threatened".*State of the Media*. (2013). Web. Fabruary 18th 2015. http://www.stateofthemedia.org/2013/newspapers-stabilizing-but-still-threatened/.

[2] Tal Galili. "Hierarchical cluster analysis on famous data sets". *R-project*. July 30th 2015. Web. August 15th 2015. https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html.

[3] Martin Maechler. "Agglomerative nesting". Seminar für Statistik, ETH Zurich. Web. August 15th 2015. https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/agnes.html

[4] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method : Which algorithms implement ward's creterion?. "Journal of Classification", 2014, 31:274-295.

[5] Hossein Arsham et Stephen P. Shao Jr. Seasonal and cyclic forecasting for the small firm. University of Baltimore, "American Journal of Small Business", N° 4, 1985, 12 Pages.

[6] J. Scott Armstrong, Roderick J. Brodie and Shelby H. McIntyre. Forecasting methods for marketing : Review of empirical research. "International Journal of Forecasting", N°3, 1987, 23 pages.

[7] Michael J. A. Berry et Gordon S. Linoff. Data mining techniques. For marketing, sales and customer relationship management. 3rd Edition, Indiana, Wiley Publishing Inc., 2011, 614 pages.