Innovation and Marketing Analytics Project: Artes s.r.l.

Bernasconi Alessia, Bottani Sophie,
Chiarini Emanuele, Giannelli Enrico

May 2022

1 Introduction

Artes S.r.l. is an Italian family-owned company founded in 1973 in Arcisate (VA). Over the past 50 years, it has grown to be one of the leaders in packaging solutions in Italy, employing more than 100 people. Its core business consists in the production of adhesive labels and stickers. Its customer portfolio includes Italian and multinational companies belonging to many different sectors, with agri-food, chemicals, and cosmetics being the most important ones.

The production of adhesive labels consists in printing different graphics onto an adhesive support by using specific technologies (for instance, digital/UV inkjet, UV flexography, UV offset, rotary screen printing). Then, labels may be embellished by coating, varnishing, or lamination.

The process requires two main categories of raw materials: the support, and the ink. Supports include different types of papers, such as coated, thermal, and vellum, as well as plastics (polyethylene, polyethylene terephthalate, polypropylene). Ink is specific to the printing technology that a product may require: for more traditional processes such as flexography, a skilled worker manually mixes Pantone colors to achieve the desired nuances, while for fully automated processes (e.g. UV inkjet), it suffices to instruct the inline printing system via a graphic file.

Over the past 15 years, the adhesive labels sector has grown exponentially, with very limited consolidation. Moreover, such growth did not ignite an increase in prices of raw materials, which largely stayed constant, or even decreased.

The typical business model adopted by industry players consisted in “on demand” production. The companies would wait for customers to place orders before purchasing raw materials from suppliers. This strategy was made possible by the very short average time of delivery ensured by paper suppliers, who have historically been able to meet their customers’ needs in 2 to 5 business days. Thus, inefficient market players were able to survive with little to no raw materials inventory, as they could wait for certainty of purchase by customers before investing in supplies. This model has turned into a vicious circle in which customers are not incentivized to carefully plan their purchases in advance, although labels are non-perishable goods requiring limited storage space.

However, the approach started showing its limits when paper supply chains were suddenly impacted by a few severe shocks.

The first wave of disruption was brought about by the Covid-19 pandemic in late 2020-early 2021. In this case, the shortage mainly referred to plastics (PET-PE-PP-PVC), which are largely purchased from China. More turmoil for the sector came in early 2022: on January 1st, the employees of one of the prime global producers of paper, a Finnish multinational company (XYZ), declared a strike which is yet to come to an end, having hit its 100th day on April 10th, 2022. The impact of this strike was particularly extensive throughout Europe, as XYZ is the biggest supplier of adhesive supports, with revenues of €1.6B.

The repercussions on the sector have been severe. Firstly, the main consequence of the shortage was an abrupt increase in prices for adhesive materials, which made honoring long-standing agreements with customers almost impossible. Secondly, the “on demand” strategy has become unfeasible due to scarcity of raw materials, which dramatically extended delivery periods to multiple months. Thus, production has stalled, resulting in delays of deliveries to customers. These adjustments ignited dissatisfaction, which has been mitigated only by general supply chains breakdowns provoked by the pandemic and the Ukrainian crisis. Furthermore, the upstream disruption might lead to striking consequences downstream: although often overlooked by the final customers, labels are of vital importance as products cannot go on the shelf without them.

In this context, the main concern for Artes is not to lose credibility and to continue to honor its agreements with established clients. Thus, the company is not always willing to transfer the repeated increases in the cost of raw materials to the end customer, as it would imply adjusting prices every week, and worse, being unable to commit to a fare set just days in advance.

Therefore, Artes looked for alternative ways to cope with the shortage. When the first hints of scarcity came about in the form of extended delivery periods, Artes increased its purchases of paper thanks to its considerable cash availability. However, this strategy of “compulsive buying” to ensure continuity of production proved to be highly inefficient, as Artes had to rent a new warehouse, thus increasing its fixed costs.

In the second half of 2021, as supply chain issues were worsening, Artes started forecasting its adhesive materials needs to ensure availability, stabler prices, and quicker production than its peers.
When XYZ announced its strike, Artes needed to act again. Firstly, the company managed to reinforce its relationship with existing smaller suppliers by ensuring payment at delivery, which enlisted Artes as a preferred customer to trade with.
Then, Artes sought new suppliers, by expanding as far as Turkey and China. The downside of such a move was limited trust, a gap that can only be bridged by long lasting relationships.

Despite some positive results, it is clear how Artes has to fully adapt its business to mitigate uncertainty, including updating its operations to stay competitive. Most importantly, the company does not have the organizational capabilities to effectively forecast its adhesive materials demand. Indeed, its purchasing department simply relies on current inventories, experience, and "gut feeling" for its planning, and more complex orders requiring specific raw materials are still processed "on demand". Thus, there is great room for improvement in planning by exploiting the data that it is able to collect from its two warehouses and its customer orders. The aim of our project is to help the company set up a data-driven planning system to better forecast its input needs and mitigate uncertainty.

2 Data Preparation

Artes provided us with two different data-sets. The first one was extracted from their business intelligence software, and includes records of all production between 2018 and 2021. It contains the following variables:

- **Log 1**: an internal order logging serial.
- **Cd Materiale**: the code of the material used to produce the adhesive labels.
- **Data Prod**: the date in which the order was produced.
- **Qta Ordin 1**: the quantity of material needed to fulfill the client's order.

It shall be noted that quantities are measured in meters, and incorporate both the different format that labels may have, and the wasted material associated to production. In fact when moving on to a new label variant, although to different extents, the printing machines all require some switch time, which implies some production scrap.
The second data-set is a much simpler one, containing Artes' professionals' input forecasts, in the form of an Excel worksheet.

We first prepare and clean our data. We noticed that some rows contain only the total quantity of a product produced over a certain period of time. We remove these rows since they are not useful in our implementation and provide information that can be easily obtained by summing all of the quantities ordered over the same time-frame. Then, we convert the date column to pandas' Datetime type to make the manipulation of time dependent variables easier and more efficient. We also create two separate columns, one for the month and one for the year of the order, as we will consider our problem from a monthly perspective.
In the data provided, the type of material is recorded only once for all orders of that same product. Thus, we repeat the correct material code for each order. Moreover, we convert the material names to the standard codes used for supply orders instead of the commercial names. During our data preparation efforts, we have repeatedly interacted with Artes' professionals to gain a better understanding of the information at our disposal. Their insight allowed us to deal with two issues we identified along the way:

- Some records reported that only one meter of a product was requested. We removed these observations as they represent samples sent to clients;
- Some orders reported quantities larger than 10,000,000. We disregarded such records as they were incorrectly registered by the BI software, since they did not appear in Artes' internal orders archive.

To assess how many units for each material the company needs to smoothly carry out production every month, we group our data in the following manner:

- By the type of material used to produce the labels;

- By month and year of the order.

Thus, we have scrapped these transformation. Lastly, as Prophet requires a year-month-day format, we join the year and month columns, and add as day the first of the month.
Note that we tried to log-scale or sqrt-scale our data to reduce the impact of high values on results, but this led to worse performance. Our final data has a total of 384 observations.

3 The model

The goal of our analysis is to effectively predict the number of meters of each material that Artes will need to serve its customers each month. The company already makes some rough estimates manually, and merely based on intuition and "gut feeling". However, if Artes can forecast the correct amount it needs to purchase, it can decrease its costs and be more efficient in the whole planning process. We choose to implement this predictive model using Prophet.

3.1 Prophet Introduction

Prophet is a package designed by Facebook's data science team whose purpose is time-series forecasting, with a special attention to business time-series. Its main advantages consist in being faster and more accurate compared to other algorithms, especially when operating with data that is incomplete, exhibiting changes in trend, and presenting outliers.

In our analysis, we first considered other models, such as auto-ARIMA, *ets*, which focuses on exponential average models, Recurrent Neural Networks (RNNs), and Kalman Filters. We discarded RNNs as we did not have access to enough data for their training and might incur in overfitting problems. Automatic ARIMA forecasts often fail to capture seasonality, while the performance of exponential averaging models was considerably inferior to that of Prophet. Lastly, for what concerns Kalman Filters, one of their main characteristics is not fully convincing in this setting. Indeed, we expect that limiting dependence of the next observation only on its previous one would not be suitable in this context, as we expect seasonality to play a big role. A potential way to avoid this issue would be to consider a given month as the current state (e.g. January 2021), and use it to predict the same month of the following year (e.g. January 2022). However, this may lead to a problematic factorisation of the trend. The latter, together with the lack of enough data induced us to progress with Prophet.

3.2 Technical Overview

At its core, Prophet is an additive model composed by the sum of three functions: a growth term $g(t)$, a seasonality term $s(t)$, and $h(t)$, a term that encapsulates idiosyncratic moments that may alter the time-series (e.g. Christmas), together with an error term $\epsilon(t)$:

$$y(t) = s(t) + g(t) + h(t) + \epsilon(t)$$

Prophet is inspired by generalized additive models (GAMs), a class of models where the response variable depends linearly on non-linear smoothers applied to the regressors. Indeed, Prophet computes forecasts through a curve-fitting exercise, without explicitly accounting for the temporal dependence of the data. This approach has several benefits compared to models such as ARIMA, including increased flexibility and easier interpretation of parameters. In the next two paragraphs, we will briefly focus on the details of the growth and seasonality components.

3.3 Trend/growth model

Concerning growth forecasting, Prophet uses a model able to change its growth rate over time. This is especially useful in the adhesive label sector, where recent breakthroughs of products such as compostable labels are boosting demand thanks to the rise of ESG standards across industries.
Suppose that we have S changepoints at times s_j, $j = 1, ..., S$. Then, we can specify a vector of adjustments

$\delta \in \mathbb{R}^S$, where δ_j is the change in rate at time s_j. The rate at any time t can be decomposed as the initial rate k plus all the adjustments until that point, i.e., $k + \sum_{j:t>s_j} \delta_j$. To represent it more cleanly, we can define a vector $a(t) \in \{0,1\}^S$ such that:

$$\begin{cases} 1, & if \quad t \geq s_j \\ 0, & elsewhere \end{cases} . \tag{1}$$

The rate at time t is therefore $k + a(t)^T\delta$. The piece-wise linear growth model is then:

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\phi)$$

where m is the offset parameter, and ϕ is the adjustment needed for the segments to connect. Alternatively, for trends that exhibit saturating growth, Prophet uses a piece-wise logistic rate of growth, but we did not implement it in our analysis of Artes. The selection of the changepoints can be either done manually or, as in our case, automatically. This is achieved through a Bayesian approach, by specifying a large number of changepoints and using the prior $\delta_j \sim \text{Laplace}(0, \tau)$, with τ controlling the flexibility of the model.
In the case of Artes, we can confidently assume both a time-varying capacity and growth rate. In fact, over the period under study, the company has been able to progressively challenge its growth ceiling, and breakthroughs of technology and new products have sustained, in some cases, a non-constant growth rate.

3.4 Seasonality

Business time-series often present seasonality as a result of human behaviours. In the case of Artes, we may expect that, during the pre-Christmas season, sales of various products such as food and cosmetics will tend to increase, leading to a boost in the demand for Artes' labels. To fit these effects, we need periodic functions of t. Prophet relies on the Fourier series, a set of mathematical functions able to approximate every curve.

Let P be the regular period that we believe characterizes the series. Then, we can approximate the seasonal effects as follows:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right)$$

The model requires the estimation of $2N$ parameters $\beta = [a_1, b_1, ..., a_N, b_N]^T$. Prophet constructs a matrix of seasonality vectors for each value of t. The following example considers monthly seasonality and $N = 5$:

$$X(t) = \left[\cos\left(\frac{2\pi(1)t}{P}\right), ..., \sin\left(\frac{2\pi(5)nt}{P}\right)\right]$$

Therefore, the seasonal component is : $s(t) = X(t)\beta$, with $\beta \sim \text{Normal}(0, \sigma^2)$, imposing a smoothing prior on seasonality.

For what concerns instead model fitting, Prophet uses Stan, a probabilistic programming language often used for Bayesian statistical inference. In particular, Prophet focuses on Stan's L-BFGS algorithm to find a maximum a posteriori estimate.

3.5 KPIs

Our analysis will focus on three KPIs: the Mean Absolute Error Percentage (MAE%), the Root Mean Squared Error Percentage (RMSE%), and the Relative Standard Deviation ($SD\%$). All metrics are expressed in percentage. For model j, we define them as follows:

$$MAE\%_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{|y_{ij} - \hat{y}_{ij}|}{\bar{y}_j}$$

$$RMSE\%_j = \frac{\sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_{ij} - \hat{y}_{ij}\right)^2}}{\bar{y}_j}$$

$$SD\%_j = \frac{\sqrt{\frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(\hat{y}_{ij} - \bar{\hat{y}}_j\right)^2}}{\bar{\hat{y}}_j}$$

where y_{ij} is the real value, $\hat{y}_{ij}$ is the predicted value, $\bar{y}_j$ is the average value of units ordered for material j, $\bar{\hat{y}}_j$ is the average predicted value of units ordered for material j, and n_j is the number of observations, which in our case are the number of months considered.

We consider the MAE%, RSME%, and SD% instead of the MAE, RMSE, or SD to make the values independent of the scale of the different materials' orders and, thus, comparable between models.

It is also important to note that we focus on both MAE% and RMSE% as they provide us with different kind of insights. Indeed, because of their mathematical definition, minimizing the MAE% will target the demand median, whereas minimizing the RMSE% will target the demand mean. As the median and the mean do not coincide, the two results will differ.

In particular, in our analysis, the median tends to be lower than the mean: on average, models fitted using the MAE% will, thus, undershoot demand for inputs. As such, using the MAE% for hyperparameters' selection and as primary metric is consistent with the company's request to favor slight underestimation of quantities rather than overstocking, in order to contain storage and handling costs. Nevertheless, we will still present the results in terms of the RMSE%, as it is more sensible to the presence of large errors and we're interested in investigating them.

Lastly, we present SD% as it grants us insight on how much variability is incorporated in our predictions. However, it is a metric that is not well suited for fitting a model or selecting hyperparameters. In fact, we would unfairly penalize prediction far from the mean even when they are the most plausible according to the estimated model. Besides, the main advantage of Prophet's flexibility resides in being able to well approximate the time series. This would be negated by choosing hyperparameters based on minimizing variance, as they would push the predictions close to a straight line.

3.6 Cross Validation and Data Partitioning

Before fitting the model, we have to select values for Prophet's hyperparameters. To fit the values of changepoint-prior-scale, seasonality-prior-scale, and changepoint-range, we will perform a data-driven validation procedure. These three parameters are fundamental: the first determines how much the trend is allowed to variate at the trend change points, the second controls the variability of the seasonality, and the third determines the portion of the data where the trend is allowed to change. To perform this procedure we divide our dataset as follows:

- Train set: all observation before January 2021.
- Validation set: observations between January 2021 and June 2021. This set is used to find the optimal values for the hyperparameters. We select those that lead to the lowest MAE%.
- New-train set: observations before July 2021. After validation, we train our model on all datapoints before July 2021 using the selected best hyperparameters.
- Test set: all observations from July 2021 onwards. It is used to test the performance of our model.

It is important to note that as we are dealing with a time-series, we cannot split our data into the different sets randomly, as it would break the temporal dependence and create look-ahead bias. Thus, we must carefully consider the chronological order of the data: we create the different sets based on when the order was placed and not randomly.

We repeat this process for the eight different types of materials the company uses, namely coated paper (PAT), vellum (VELLUM), thermal paper (TH), white polypropylene (PP W), clear polypropylene (PP C), white polyethylene (PE W), white polyethylene terephthalate (PET W), and supports for digital/UV inkjet printing (DIG). In this way, we fit eight different models, one for each material, with the respective best hyperparameters.

4 Results

4.1 Results for the full training set model

After fitting the optimal models on the whole training set, we test their performance on the test set. Their results can be seen in table 1.

When analysing the MAE%, we can see that we get great results for orders made in PAT, PP W, or VELLUM, with a MAE% ranging from 5% to 16%, we get satisfactory scores for orders in PE C, TH, and for DIG, with a MAE% varying between 27% and 37%, and poorer results for PET W and PP C, with a MAE% of 49% and 84% respectively.
If we inspect the RMSE%, as expected, we get slightly higher results, but the magnitude of the difference is not concerning, with the exception of PET W. In fact, its RMSE% reaches 109%, up 25% with respect to its MAE%. Such difference likely signals the presence of large errors, to whom RMSE% assigns higher weight.
Lastly, we look at the SD%. Models that performed worse according to MAE% and RMSE% tend to confirm this inefficiency, showing higher variability and a consequent lower reliability of predictions. Note, though, that in general, models that show worse fits are those whose order values present high variability, which may reflect into more volatile predictions.

We can confirm our insights by looking at Figure 1, where we plot the actual time-series versus our predictions. Indeed, the plots of PAT, PP W, and VELLUM show that the two paths almost perfectly cross in the test data portion, certifying that we were able to model their characteristics very well. Instead, by looking at the plots of PE C, TH and DIG, we can observe that while the overall trend has been predicted correctly, the model has failed to pick up some of the variability and has preferred to smooth it out. The resulting performance is weaker than that of previous materials, but we still achieve satisfactory predictions. In the plot for PP C, we see that the predictions are a lot less precise and that our model fails to pick up some characteristics of the path, as the relative error suggested.
Lastly, we focus on PET W: our model preferred a smoother path, although not as smooth as those of other models, which to a certain extent is able to mimic the actual time-series fairly well. However, PET W showed the highest RMSE%, so we expected large errors. In fact, we can see how in the first observation, there is quite a big difference between the predicted and actual use of PET W, which is likely to be driving most of the difference between RMSE% and MAE%. Furthermore, if we compare average quantities over the period, we can see how PET W has one of the lowest averages among all materials, and in fact presents a decreasing trend. When discussing this evidence with Artes' professionals, we learnt that they are progressively trying to cutback on production of PET labels, in order to reduce the environmental footprint of the company's operations. Traditionally, PET has been a preferred support as it is able to guarantee high printing quality, and vividness of colours. However, technological advances in digital printing are enabling Artes to achieve comparable quality on DIG materials, which are much more sustainable. Thus, they tend to purchase PET W only when strictly necessary, or explicitly requested by the client.
To conclude, we can confirm how the materials for which our model performs worse are those which seem to show the most erratic behaviours over the 48 months, often characterized by very sharp peaks and drops, and, in general, a variability that is not present in the other paths. Thus, we would probably require a larger data-set to get an improved fit that does not result in a smoothing of the function.

It is also interesting to analyse Figure 2, where we plot our predictions against the true values. This allows us to validate our insights. In the models that performed better (PAT, PP W and VELLUM), we see that almost all the values are very close to the 45° degree line. For PE C, TH, and DIG, this is still true, but some points start getting distant. Lastly, for PP C and PET W, we see how values are now a lot more sparse, confirming what we saw in the error rates and in the previous plots.

Lastly, we can take a look at the trend-seasonality decomposition plots. We illustrate and comment on those of PP C and TH in Figure 3, as they are very representative. Indeed, PP C presents a decreasing trend, which can also be noticed in PET W, while the growing trend displayed by TH is common to DIG, PE C, and PP W. For the moment, we ignore PAT and VELLUM as we will discuss them in the next subsection.
Firstly, we can notice how the slope of the trends are opposite: over the period under study, Artes has managed to grow its revenues year after year, so we would expect all trends of use of materials to be increas-

ing. However, two aspects shall be noted: downward oscillations are driven by an upgrade in production technologies (and particularly those involving PP C), which allowed the company to drastically decrease waste, and the use of some materials (particularly PET W) being progressively reduced in an effort to tackle down the environmental impact of Artes' production.
Secondly, both examples, as well as most other materials, have peaks in July. This is most likely due to the fact that Artes reduces its production consistently in August, inviting customers to stock up on labels in July. Moreover, in both plots demand tends to be higher in the later parts of the year with respect to earlier quarters, which might be due to stronger sales most companies enjoy in the holiday period. Such boost is likely to be driven by the agri-food sector, as Artes serves many supermarket chains, as well as many prominent exporters of Italian products, which experience booms during the Christmas season.

4.2 Results for the reduced training set model

To further test the robustness of our models, we fit them on the first 30 months of the data only (observations before June 2020), and then test their performance on the following 6 months (observations between July 2020 and December 2020).

Results are summarized in table 2. If we compare these findings to the full training set model, we obtain some interesting results. Our best model was PAT, with MAE% around 5% and RMSE% around 6%. Such errors have worsened significantly by training the model on the reduced set, with MAE% at 33% and RMSE% at 37%. On the other hand, for PP C, which was one of our worst models in the previous fit, results have improved substantially, with MAE% dropping from 49% to 35%, and RMSE% from 55% to 42%. A similar improvement can also be noticed for PE C.
Furthermore, we can see how the performance of the models for TH, PP W, and PET W has slightly worsened, while that of VELLUM and DIG remained almost constant.
If we focus on the RMSE% for PET W, which was problematic in our first fit, we can see how the new model performed almost identically to the previous one, signaling that the largest share of the variability is experienced at the beginning of the period, as the reduced training set model is still able to pick it up.
Looking at the SD%, the results are even more consistent to the ones of the previous models, confirming our hypothesis that the variability of the model was driven by the variability of the observations and thus, models that perform worse are those characterised by paths with more intrinsic variability. The only big change in this metric happens to PP W's model which increased its variability significantly, in line with the increase in errors we pointed out. Overall, almost all the models still present remarkably good results, given that they were fitted on only 30 months of data.

We can confirm these findings by looking at Figure 4. The plot for VELLUM (the best performing model now) confirms that Prophet was able to pick up the main characteristics of the trend and predict in an almost perfect way over the last 6 months. Similarly, most of the other plots show more than satisfactory results, with models being able to pick up the main features of the path and predicting well. However, in some cases, the lower amount of training data did not allow for a full characterization of the path. For instance, if we consider PAT, we see how short-term drops that were previously ignored are now driving the predictions, thus leading to bigger errors.

Similar insights can be obtained by looking at Figure 5. In most cases, values are very close to the 45° degree line, thus confirming the good fit we achieved. For example, in the case of VELLUM, all of the points are incredibly close to the line, as we expected from the previous analysis. Conversely, if we look at the plot of PAT, we can clearly see how we are under-predicting values in the new portion of data, signaling that our model is probably unable to pick up the correct trend.

As previously mentioned, the PAT model has worsened considerably when reducing the training set. We further investigate the shift in performance by looking at the trend and seasonality plot for the full and reduced models (Figure 6). It is evident how the trend changed completely at the beginning of 2020, getting an higher positive slope. Hence, the model fitted on the reduced training set was unable to incorporate this information, as its training sample ended in June 2020. This fact is probably the root cause of the worsening of the fit. On the other hand, the seasonality is very similar between the two models, suggesting that the smaller data-set managed to capture it.

Lastly, it is interesting to look at the trend and seasonality plot for VELLUM (Figure 7) to try and understand the drivers of the performance improvements. Firstly, we can observe the different trends characterising the two models: for the full training sample, an initial decreasing trend is followed by an upward sloping one in mid-2020. Conversely, for the reduced training set, the trend continues going down beyond the first half of 2020. As in the case of PAT, these findings suggest that the smaller training sample did not enable our model to pick up the correct trend. However, it shall be noted VELLUM is characterised by a very small magnitude of the slope, thus, such mistake is much less concerning than for PAT.

Once again, the seasonality is almost identical for the full and reduced training set models. Interestingly, we see that, for vellum, the biggest drop of the year happens in June, instead of August: vellum is mostly used to produce neutral, blank labels, which are then printed over by customers. Thus, such stickers are not custom-made, and can be sold to multiple clients. A prime example are the white labels listing ingredients of agri-food products. Every year, Artes schedules its production of blank labels in such a way not to bottle up its manufacturing capacity during the busiest times of the year. In fact, in June, most cosmetic companies prepare to launch their summer collections, which normally require very elaborated labels, as they put a lot of attention on packaging. Thus, Artes prefers to stock up on neutral labels in May, and direct this capacity to potential last-minute orders or changes. Along this line, although less prominent, another decrease can be seen around December.

4.3 Prediction Comparison

As previously mentioned, Artes makes manual estimations of future orders. Since we were provided with such data, we are interested in statistically assessing whether our model would have performed significantly better than the company's estimates. Thus, we compare the test error rate of our predictions over the last six months of data with the error rate of the company's estimations by implementing a simple comparison of proportions test. We will consider the company as treated when using our prediction model, and as control when using internal predictions.

Thus, what we are interested in testing is the following hypothesis: $H_0 : MAE\%_t = MAE\%_c$ vs. $H_1 : MAE\%_t < MAE\%_c$

Under H_0, we have $nMAE\%_t \sim \text{Binomial}(n, MAE\%_c)$. Therefore,

$$\text{p-value} = P(MAE\%_t < \widehat{MAE\%}_t) = P(nMAE\%_t < n\widehat{MAE\%}_t)$$

where $\widehat{MAE\%}_t$ is the observed error rate on the treated, n is the number of months we consider. We can simply use the binomial CDF to obtain this p-value. Note that in our analysis, we will use the standard significance level $\alpha = 5\%$.

The results are reported in table 3.

In general, we can notice that for all materials, our model's MAE% is lower than that of Artes' manual predictions. It is interesting to note that, while enjoying greater stability, the company's errors tend to be higher for the same materials where our model's error rate is large, with the exception of PET W. Thus, we can confirm that our error rate is higher for materials whose demand varies more erratically, as we inferred from the plots. In order to reject the null hypothesis and claim that our models' forecasts are more reliable than the company's ones, we have to analyze the p-values obtained.

We achieve our best result for PAT, as our forecasts outperform the company's by 47%, with a p-value of 0.0113. A striking improvement is also achieved for PP W (43%), TH (36%), and PE C (32%). All other models present remarkable increases in accuracy (all above 10%), with the exception of PET W, whose gain is less than 2%. Nevertheless, under our significance level α, only four out of the eight models reject the null hypothesis. A reason for this outcome is the low power of the test. In fact, the two proportions are built on a low number of observations (6). Thus, it is reasonable that we will not reject the null hypothesis unless the difference in the predictions' quality is sizeable, as in the case of PAT.

In order to get more general results and not material-specific ones, we also compared the average MAE% across all models, and that of the company's predictions. In this case, given the larger sample size, we can use the normal approximation of the proportion test. The formula for the test statistic is, thus:

$$z_t = \frac{\overline{MAE\%}_t - \overline{MAE\%}_c}{\sqrt{\overline{MAE\%}(1 - \overline{MAE\%})(\frac{2}{n_{tot}})}}$$

where

$$\overline{MAE\%} = \frac{\overline{MAE\%}_t + \overline{MAE\%}_c}{2}$$

and $\overline{MAE\%}_t$, $\overline{MAE\%}_c$ are the average MAE% made by our model, and by the company's predictions respectively, and $n_{tot} = n_{months}(n_{models})$, where n_{months} and n_{models} are the number of months and models considered respectively.By performing this test, we get a $z_t = -17.1259$ which has a $p\text{-}value < 0$.
We also computed the average treatment effect (ATE) as

$$ATE = \overline{MAE\%}_t - \overline{MAE\%}_c = -0.2612$$

Thus, we can conclude that overall our model outperforms the company's predictions by 26%, and that such difference is statistically significant.

5 Organizational implications

5.1 Representitativeness, Bias, Replicability

Our model is aimed at helping an adhesive label company in forecasting its input needs. First of all, we shall assess to what extent Artes is representative of the adhesive label sector in general, and in Italy. As previously mentioned, Artes is one of biggest label makers in Italy. Thus, within the country, there might be much smaller players with even more limited data availability, which might represent a difficulty for our model's applicability and success. On the other hand, there might be players whose technological capabilities are more advanced than Artes': they may be already equipped with well-performing planning systems, and thus, would not benefit from our suggestions as much as Artes would.
Then, we shall consider bias. Artes was willing to collaborate with us as it considered our project as an opportunity to expand its non-manufacturing technological capabilities, which are very limited, as it lacks an IT division. Other companies might be less willing to share data with us as they might prefer to internalize such an important resource.
Lastly, we consider how likely our study would be applicable to other companies. Raw materials scarcity is a very relevant issue, and has been experienced by the entire sector at European level. However, not all companies may have been affected by the shock symmetrically. By interacting with Artes' professionals, we have found out that many competitors have struggled with the same issue, and some were not able to cope due to limited beginning-of-the-year inventory. As there is homogeneity of raw materials across the sector, and most companies refer to a few big suppliers, these peers might benefit from the trial of our model and our suggestions.

5.2 Suggestions

Our analysis has led us to develop many suggestions that may help the company in succeed in these uncertain and complex times.
First of all, given this last result, we advise the company to implement our model in practice to predict its monthly paper needs. We suggest, though, that for a period of time it also keeps preparing its own predictions so as to compare the performances of the two methods and assess whether the improvement found in our results would continue to hold in the future.
Our analysis was faced with many limitations due to the data we were provided with and that is logged by the company. Firstly, within the company, data on production is collected in two ways. For customer orders, the process is fully automated. Once a label is designed, it is assigned to a production card which contains all the information needed for its manufacturing, from materials, to format and colour schemes. Then, every time Artes receives a new order for that specific label, the production card and the desired

quantity are digitally transferred to the machine in the factory, which is instructed with all the data required for production. Finally, the machine produces and collects data through a business intelligence tool, which can be retrieved from the company's internal databases. This process presents two main issues: automation may lead to mistakes in compiling data, as we witnessed in our analysis, and the BI has specific formats, such as a built-in pivot tool, which make data preparation more lengthy. Furthermore, the company might be interested in predicting both its paper and ink needs. However, the production data it currently collects do not contain information on the type of ink used. Thus, we were not able to apply our model to this specific production input. Concerning data on inventories, all materials entering and exiting the company's warehouses are reported manually. It is clear how this practice is inefficient as it is time-consuming and might result in human error. Hence, we think the company should invest in its data capabilities both by improving its software set and by hiring data-competent employees to ensure it is able to exploit as much knowledge as possible in order to support its manufacturing.

More in general, it can be argued that Artes is a traditional real economy firm, deeply rooted in its family-owned culture, which to a certain extent shielded it from many turmoils over the past 50 years. However, as the world progresses towards a fully data-driven approach, it is evident how its operations need to adapt to avoid passing by huge opportunities. We think the company would benefit from extending its non-manufacturing investments, in an effort to boost growth, optimize production processes, and create an even stronger competitive advantage. Although over the past 10 years it has consistently invested in cutting-edge production equipment, which allowed it to minimize production waste, as well as to achieve remarkable printing quality, such forward-looking approach should be applied all over its operations, from inventory forecasting, to employee management.

Lastly, we propose the management of the company to explore one further weakness we witnessed in our analysis of the company: its lack of an appropriate salesmen performance payment scheme. In fact, Artes bases the largest share of its sales on the expertise of 20 salesmen, each assigned to a specific area. Most have been with the company for more than 20 years, and have great knowledge of the sector and its operations. However, their salaries do not correctly incentivise their activities: although they earn a commission on their orders, such percentages are based on agreements signed many years ago, and are client- and label-specific. We learnt how commissions tend to be higher for long-standing customers, and lower for newer ones, in an effort to offer more competitive prices. However, such rates disincentivise effort on the agents' side, which are less likely to go out of their way to find new clients, and more likely to extract the most from well-established relationships. Due to disclosure reasons, we were only granted access to production data. Had we been able to access also the sales' data, we would have tried to create a way to adjust the salesforce salaries in a clear, data-driven way, so as to better align their incentives to the company ones.

6 Appendix

Material	MAE%	RMSE%	SD%
DIG	0.369934	0.410629	0.275715
PAT	0.052652	0.062324	0.136134
PE C	0.288166	0.335131	0.250274
PET W	0.842623	1.095224	0.806277
PP C	0.492276	0.551310	0.413767
PP W	0.152894	0.194148	0.266167
TH	0.275441	0.381333	0.163074
VELLUM	0.164430	0.226301	0.207732

Table 1: The results on the test set of the model fitted on the full training set (up to June 2021) using the parameters found by the cross-validation procedure.

Material	MAE%	RMSE%	SD%
DIG	0.366163	0.395247	0.310944
PAT	0.329899	0.372443	0.157337
PE C	0.213408	0.239596	0.219678
PET W	0.925642	1.119328	0.885972
PP C	0.353593	0.416090	0.429140
PP W	0.282158	0.335359	0.474798
TH	0.451561	0.554016	0.174740
VELLUM	0.141302	0.158172	0.205503

Table 2: The results on the last six months of 2020 of the model fitted on the reduced training set (up to June 2020) using the parameters found by the cross-validation procedure.

Material	$\widehat{MAE\%}_t$	$\widehat{MAE\%}_c$	$\widehat{MAE\%}_t - \widehat{MAE\%}_c$	p-value
DIG	0.369934	0.510935	-0.141000	0.3235
PAT	0.052652	0.526499	-0.473846	0.0113
PE C	0.288166	0.612664	-0.324497	0.0354
PET W	0.842623	0.859918	-0.017294	0.5957
PP C	0.492276	0.675232	-0.182956	0.0920
PP W	0.152894	0.591621	-0.438726	0.0046
TH	0.275441	0.637834	-0.362392	0.0261
VELLUM	0.164430	0.313786	-0.149356	0.1044

Table 3: A comparison between our model's error rates and the company's ones on the last six months of 2021 (our test set). The p-values are obtained using binomial proportions tests.

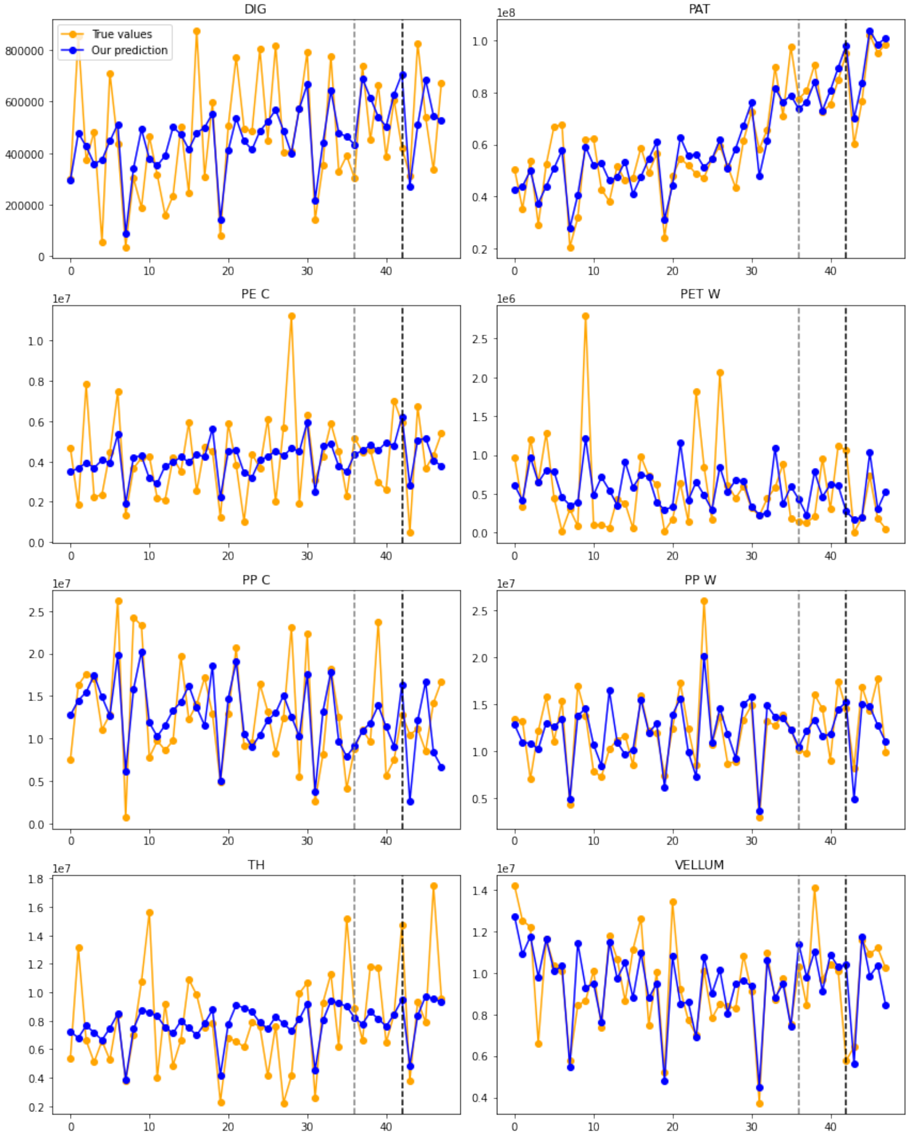

Figure 1: The actual time series (in orange) vs the predictions made by our model fitted with data up until June 2021 (in blue) for each of the 8 materials. The grey (to the left) and black dotted lines (to the right) indicate respectively the beginning of the validation set and the one of the test set.

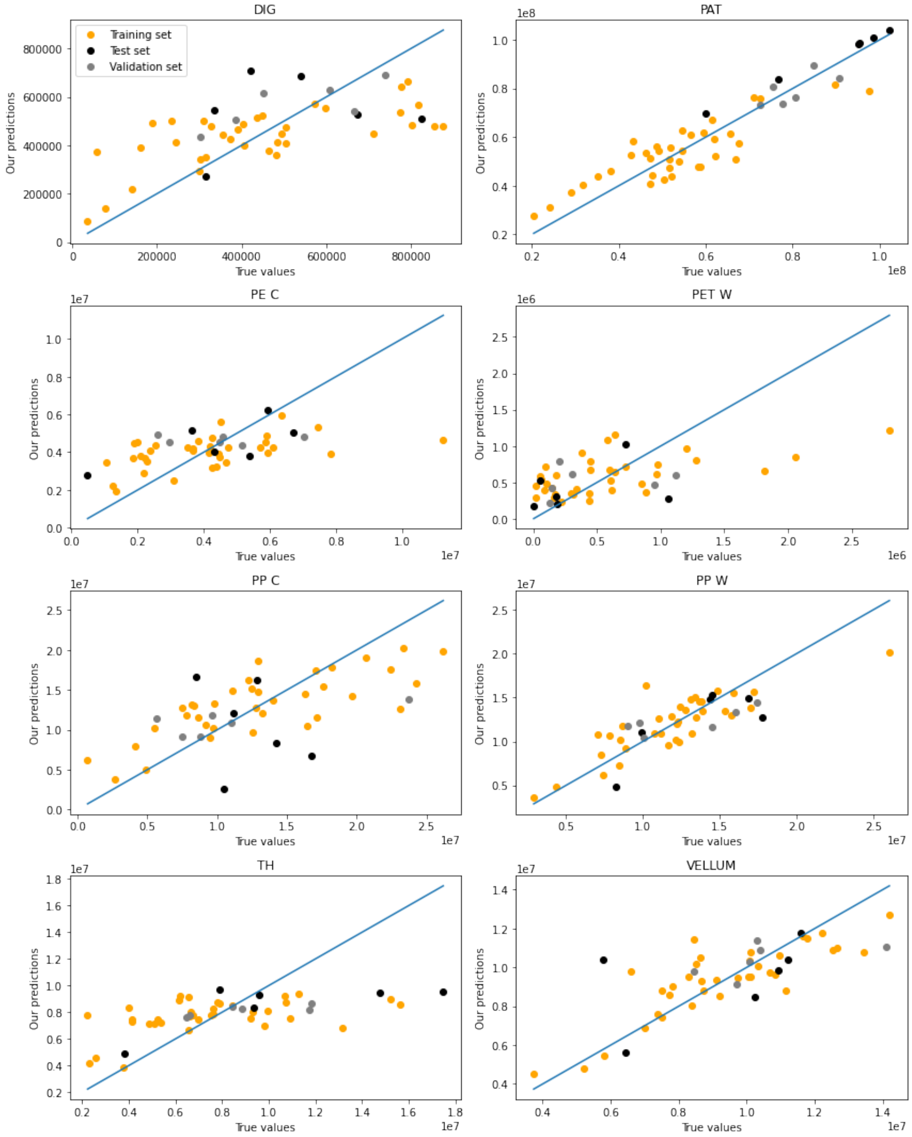

Figure 2: A plot of our model's (trained on the full training set) predictions against the true values of the time series.

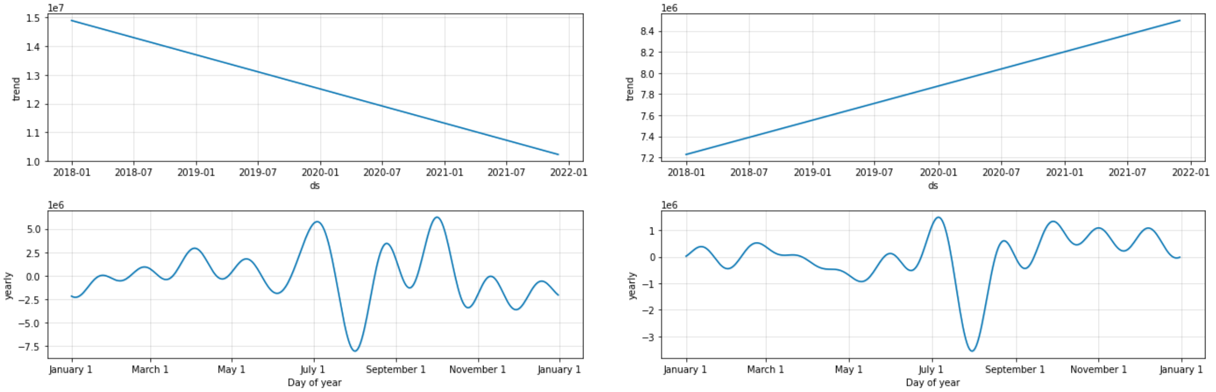

Figure 3: Trend and seasonality plots for PP C (left) and TH model (right) when fitted on the whole training set.

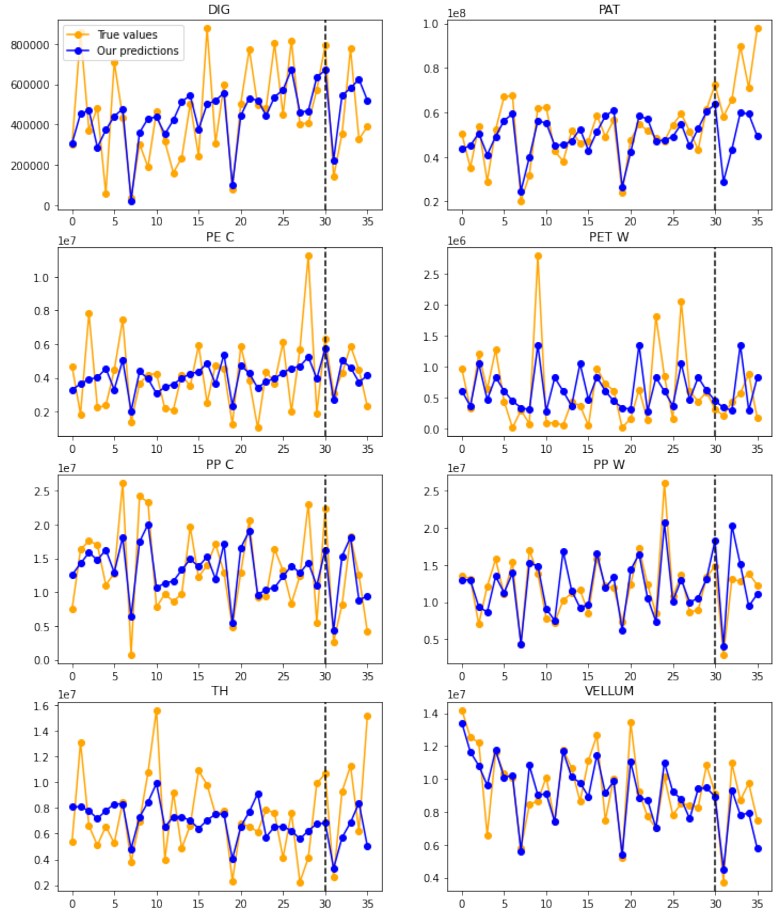

Figure 4: The first part of the actual time series (in orange) vs the predictions made by our model fitted with data up until June 2020 (in blue) for each of the 8 materials. The black dotted line indicates the beginning of the test set.

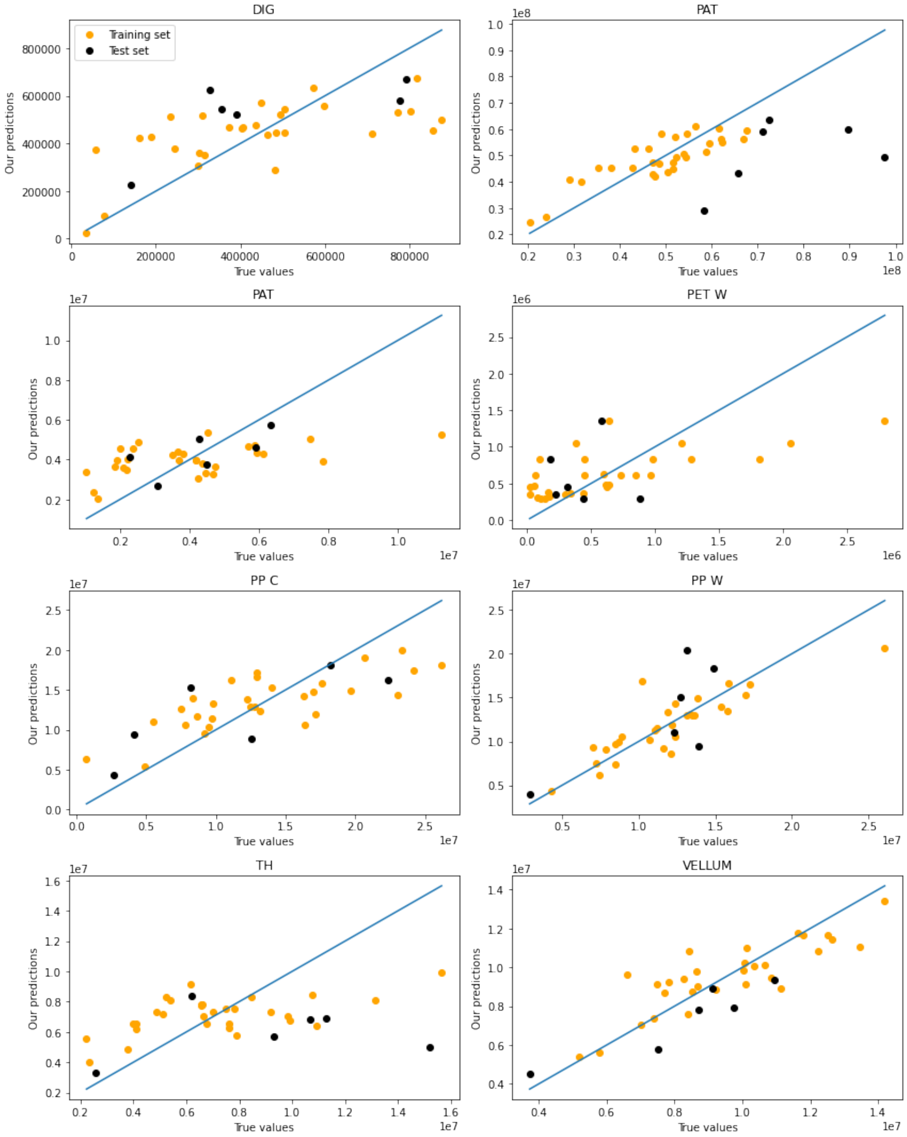

Figure 5: A plot of our model's (trained on the reduced training set) predictions against the true values of the time series.

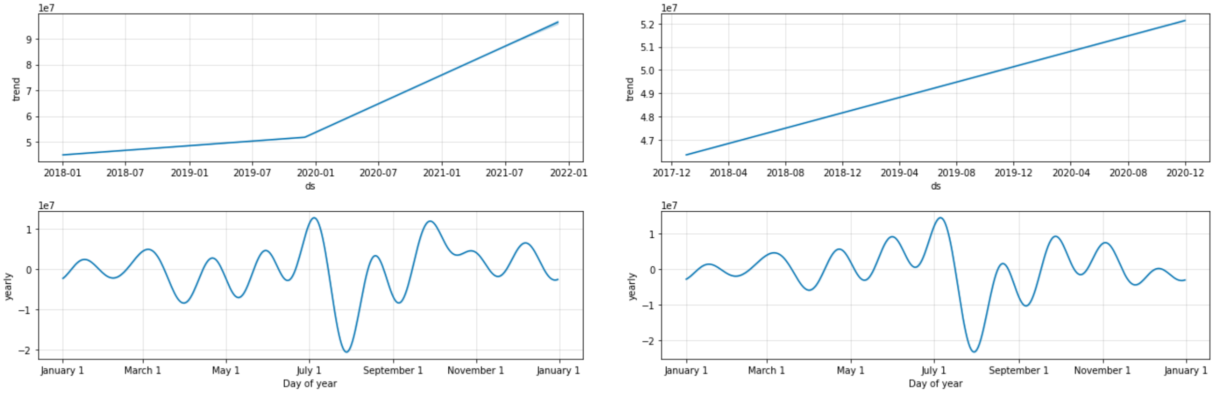

Figure 6: Trend and seasonality plots for PAT's model when fitted on the whole training set (left) and on the reduced one (right).

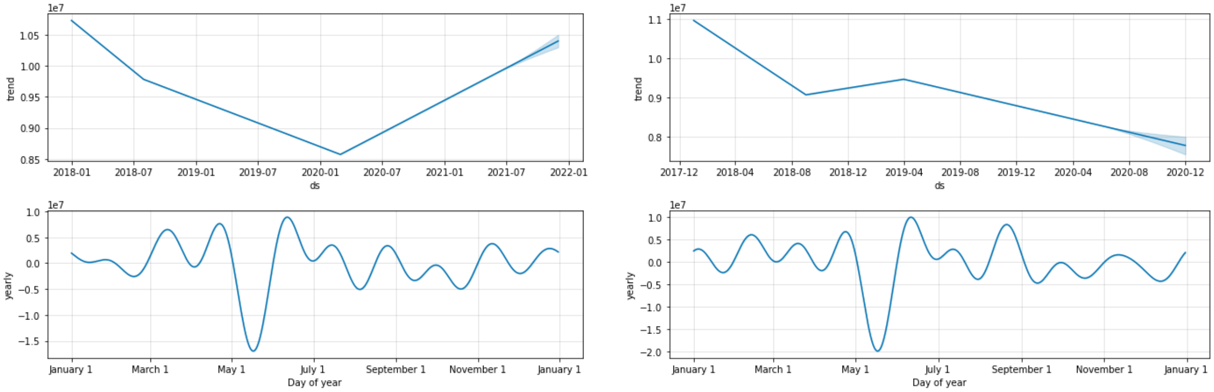

Figure 7: Trend and seasonality plots for VELLUM's model when fitted on the whole training set (left) and on the reduced one (right).