Degree Project in Computer Science

Second cycle, 30 credits

Forecasting post COVID-19

How to improve forecasting models' performance when training data has been affected by exceptional events like COVID-19 pandemic?

LINA SHREBATI

Stockholm, Sverige Klicka här för att ange aktuellt årtal

Forecasting post COVID-19

How to improve forecasting models' performance when training data has been affected by exceptional events like COVID-19 pandemic?

LINA SHREBATI

Master's Programme, Machine Learning, 120 credits
Date: September 21, 2023

Supervisor: Anubhab Ghosh
Examiner: Saikat Chatterjee
School of Electrical Engineering and Computer Science
Host company: Artefact
Swedish title: Prognos efter covid-19
Swedish subtitle: Hur kan man förbättra prognosmodellens prestanda när träningsdata har påverkats av exceptionella händelser som COVID-19-krisen?

Abstract

Almost every company around the world were affected by the COVID-19 crisis and the government measures that were taken to slow the spread of the virus. The impact the crisis had on the economy caused the appearance of anomalies in the data collected by companies : such as abnormal trend, seasonality etc. Traditional methods of forecasting were then questioned when trying to predict business indicators such as sales in a post COVID-19 world, as we saw performance like forecast accuracy decreased. So how can data scientists increase the performance of their forecasting models in a post COVID-19 world knowing that the training data contains COVID-19, an event never observed before? What methods can be used to overcome this problem? The goal of this project was to provide a guideline for dealing with COVID-19 data points for forecasters. We first dedicated this thesis to data analysis and finding a clear methodology to better understand and quantify the impact of COVID-19 crisis on business indicators. Then, we compared multiple methods to overcome the forecasting issues that are faced when training datasets influenced by the phenomenon of COVID-19 and improved forecast accuracy and reduce bias. Each method had its pros and cons. Among the methods changing the training data, imputation is the easiest method and can give very good results. Multiplicative coefficients also can be used, and give also good results. Finally, optimal transport was tested as an alternative to the two first methods. This method changes less the original the time series compared to imputation. Among methods consisting in adding external features to the model, a boolean feature is the most simple way to flag a COVID-19 period and works surprisingly well. Adding more complex features describing COVID-19 impact on the time series is challenging since we need to find a feature that describes well the phenomenon and be able to use another model to predict its future values if we want to use it for our first model. Adding Google mobility features to the model as external regressors seem to increase the most forecast accuracy, but its performance depends on how well we can estimate their future values. This applies also to stringency index, but predicting stringency index future values is even harder as we are trying to estimate government measures. However, with the Stringency index we can simulate scenarios if we make a hypothesis on future government measures: we can estimate COVID-19 impact on the time series in a worst case scenario with lockdowns by setting the Stringency index high for instance.

Keywords

Time Series, Forecasting, COVID-19, Data processing, Optimal transport

Sammanfattning

Nästan alla företag runt om i världen drabbades av covid-19-krisen och de statliga åtgärder som har vidtagits för att bromsa spridningen av viruset. Krisens inverkan på ekonomin orsakade uppkomsten av anomalier i data som samlats in av företag: onormal trend, säsongsvariationer ... etc. Traditionella metoder för prognoser ifrågasattes sedan när man försökte förutsäga affärsindikatorer som försäljning i en värld efter covid-19, eftersom vi såg att prestanda som prognosnoggrannhet minskade. Så hur kan dataforskare öka prestandan för sina prognosmodeller i en värld efter covid-19 med vetskapen om att träningsdata innehåller covid-19, en händelse som aldrig tidigare observerats? Vilka metoder kan användas för att övervinna detta problem? Målet med detta projekt var att ge en riktlinje för hantering av covid-19-datapunkter för prognosmakare. Vi dedikerade först denna avhandling till dataanalys och att hitta en tydlig metod för att bättre förstå och kvantifiera effekten av covid-19-krisen på affärsindikatorer. Sedan jämförde vi flera metoder för att övervinna problemet med den COVID-19-påverkade träningsdatauppsättningen och förbättrad prognosnoggrannhet och minskad bias. Varje metod hade sina för- och nackdelar. Bland metoderna för att ändra träningsdata är imputering den enklaste metoden och kan ge mycket goda resultat. Multiplikativa koefficienter kan också användas och ger också bra resultat. Slutligen undersöktes en ny metod: optimal transport, och kan vara ett alternativ till imputering. Med denna metod är den ursprungliga formen på tidsseriekurvan lite mer bevarad, så viss information i originaldata är fortfarande användbar för modellen. Bland de externa funktioner som lagts till i modellen är den booleska funktionen det enklaste sättet att flagga en covid-19-period och fungerar förvånansvärt bra. Googles mobilitetsfunktioner är de externa regressorer som verkar öka mest prognosnoggrannhet, men det beror på hur väl vi kan uppskatta deras framtida värden. Detta gäller även stringensindex, men ännu svårare då vi försöker skatta statliga åtgärder. Stringensindex kan användas för att simulera scenarier (värsta scenario med låsningar, bästa fall där allt är öppet).

Nyckelord

Tidsserier, Prognoser, COVID-19, Databehandling, Optimal transport

Résumé

La crise de la COVID-19 et les mesures gouvernementales qui ont été prises pour ralentir la propagation du virus ont touché presque toutes les entreprises du monde. L'impact de la crise sur l'économie a provoqué l'apparition d'anomalies dans les données collectées par les entreprises : tendance anormale, saisonnalité modifiée...etc. Les méthodes traditionnelles de prévision ont ensuite été remises en question lorsque nous avons essayé de prédire des indicateurs commerciaux tels que les ventes dans un monde post-COVID-19, car nous avons vu des performances telles que la précision des prévisions diminuer. Alors, comment les Data Scientists peuvent-ils augmenter les performances de leurs modèles de prévision dans un monde post COVID-19 sachant que les données d'entraînement ont été affectées par la COVID-19, un événement jamais observé auparavant ? Quelles méthodes peuvent être utilisées pour surmonter ce problème ? L'objectif de ce projet était de fournir une ligne directrice pour traiter les données impactées par la COVID-19 pour les Data Scientists. Nous avons d'abord dédié cette thèse à l'analyse des données et à la recherche d'une méthodologie claire pour mieux comprendre et quantifier l'impact de la crise du COVID-19 sur les indicateurs commerciaux. Ensuite, nous avons comparé plusieurs méthodes pour surmonter le problème des données d'entraînement affecté par la COVID-19 et améliorer la précision des prévisions et réduire les biais. Chaque méthode avait ses avantages et ses inconvénients. Parmi les méthodes modifiant les données d'apprentissage, l'imputation est la méthode la plus simple et peut donner de très bons résultats. Des coefficients multiplicatifs peuvent également être utilisés et donnent également de bons résultats. Enfin, une nouvelle méthode a été explorée : le transport optimal, et peut être une alternative à l'imputation. Avec cette méthode, la forme d'origine de la courbe de la série temporelle est un peu mieux préservée, de sorte que certaines informations contenues dans les données d'origine sont toujours utiles pour le modèle. Parmi les données externes ajoutées au modèle, la variable booléenne est le moyen le plus simple de signaler une période COVID-19 et fonctionne étonnamment bien. Les données de mobilité de Google sont les régresseurs externes qui semblent augmenter le plus la précision des prévisions, mais cela dépend de la façon dont nous pouvons estimer leurs valeurs futures. Cela s'applique également à l'indice de rigueur, mais celui ci est encore plus difficile car nous essayons d'estimer cette fois ci des mesures gouvernementales. L'indice de rigueur peut être utilisé pour simuler des scénarios (pire scénario avec confinement, meilleur scénario où tout est ouvert).

Mots clés

Séries temporelles, Prévision, COVID-19, Traitement de l'information, Transport optimal

Acknowledgments

I would first like to thank my colleagues at Artefact. First, my supervisor at Artefact, Antoine Aubier, for having supervised this project and having provided me with the data I was able to work on. Antoine was of immense help throughout the project. I would also like to thank Maxime Lutel, who also helped me on this project and encouraged me to work on different methods including optimal transport. Without them, this project would not have had the same results. Also, at KTH I had the pleasure to work on my report with the help of my supervisor Anubhab Ghosh. Finally I would like to thank my examiner at KTH, Saikat Chatterjee, for trusting me in accepting this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The COVID-19 health crisis and the government measures that have been taken to slow the spread of the virus have had a significant impact on most sectors of the economy: fewer physical sales during lockdown, increased activity in call centers following the announcement of the lockdown, increased purchase of certain products, etc.

These phenomena were not predicted by most demand prediction models due to the exceptional nature of this health crisis, and therefore demand could not be anticipated causing stockouts or overstocks for some stores.

In addition to the unpredictability of this crisis, the change and slowdown in activity during periods when government measures were strongest (containment, border closures, ban on public gatherings, etc.) caused the appearance of atypical observations in the data: absence, decrease or increase in the signal, change in seasonality, etc. The performance of models trained on data containing these abnormalities at predicting post COVID-19 crisis (e.g. demand) has decreased as a result.

The challenge that many data scientists encounter today subsequently is to find methods to process the historical data that has been affected by the slowdown in economic activity due to the health crisis or to find models for the pandemic to feed to demand forecasting models in order to increase the performance of the latter.

1.1 Background

The COVID-19 pandemic, which caused the largest global recession since the Great Depression in 1929, has affected businesses and industries across the globe which is reflected in their underlying data.

In France for example, the following measures were taken by the government to limit the spread of the virus:

- Feb 29th, 2020 : Cancel public events and Restrictions on gatherings (France bans large indoor gatherings with more than 5000 attendees)
- Mar 8th, 2020 : Restrictions on gatherings (Gatherings of over 1000 people banned across France)
- Mar 13th, 2020 : Restrictions on gatherings (France bans gatherings of more than 100 people)
- Mar 14th, 2020 : Restrictions on internal movement - recommend not to travel between regions/cities
- Mar 16th, 2020 : School closing - require closing all levels and Close public transport - recommend closing (or significantly reduce volume/route/means of transport available)
- Mar 17th, 2020 : Workplace closing - require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors) and Stay at home requirements - require not leaving house with exceptions for daily exercise, grocery shopping, and 'essential' trips

Lockdown was held until May 11th, 2020.

As a consequence, of those measures, the economy was heavily affected. The 2020 stock market crash began on 20 February 2020, although the economic aspects of the COVID-19 recession began to materialise in late 2019.

France has been hit hard by the pandemic, with two months of 'strict lockdown' imposed before mid-year. On 8 April 2020, the Bank of France declared that the French economy was in recession, shrinking by 6 percent in the first quarter of 2020.

At the end of the second trimester of 2020, several companies began staff cuts: Nokia (1233 jobs), Renault (4600 jobs), Air France (7580 jobs), Airbus (5000 jobs), Derichebourg (700 jobs), TUI France (583 jobs) and NextRadio TV (330–380 jobs).

In fact, those companies often anticipated the decrease of demand that they were going to endure. The below graph illustrates the total volume sales of a giant retailer, year 2020 being highlighted in red.

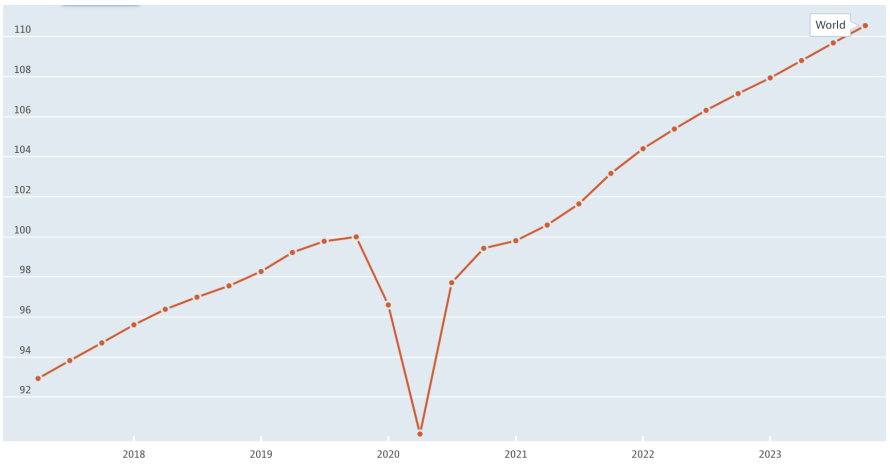

Figure 1.1: World GDP evolution (in level percentage before crisis)

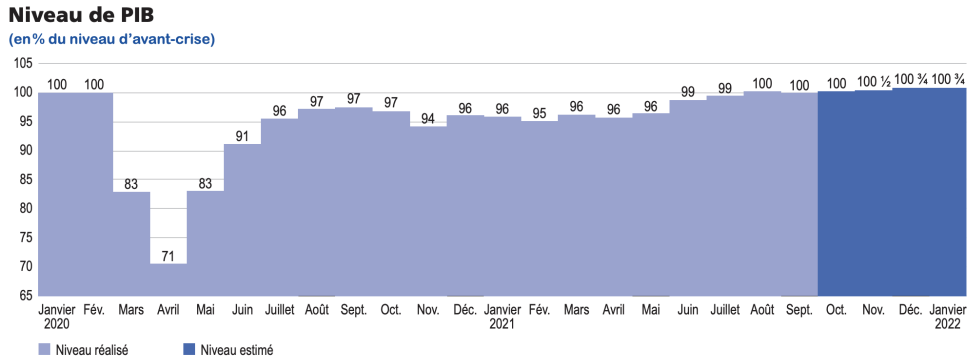

Figure 1.2: France GDP evolution (in level percentage before crisis)

We can observe a smaller peak July 2020 compared to previous peaks in 2019 and 2018. Also, the same effect on demand was observed for a large luxury group.

For call centers, a more subtle effect was observed, and is harder to detect by observing directly the total volume of calls per day.

1.2 Problem

The impact to each company largely varies depending on the industry where it operates and how it has been possible for this company to respond to the ongoing crisis

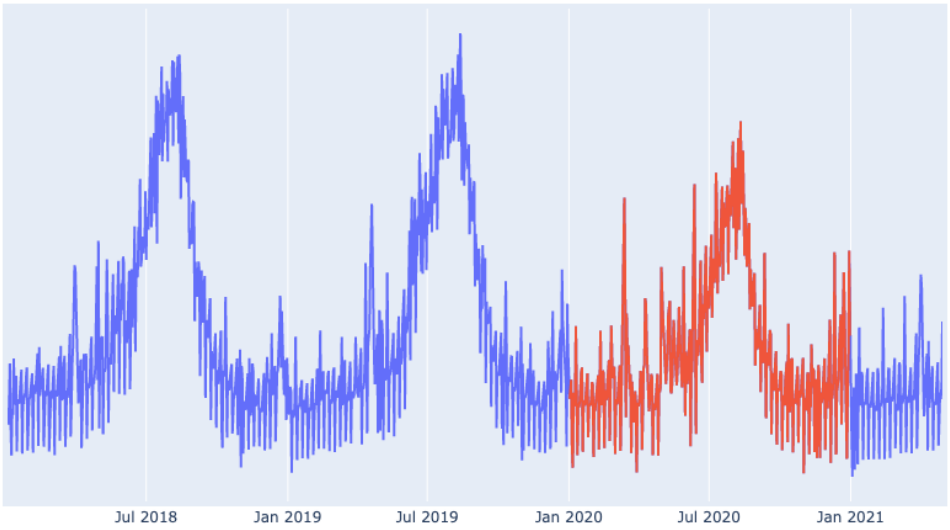

Figure 1.3: Daily volume sales for a giant retailer

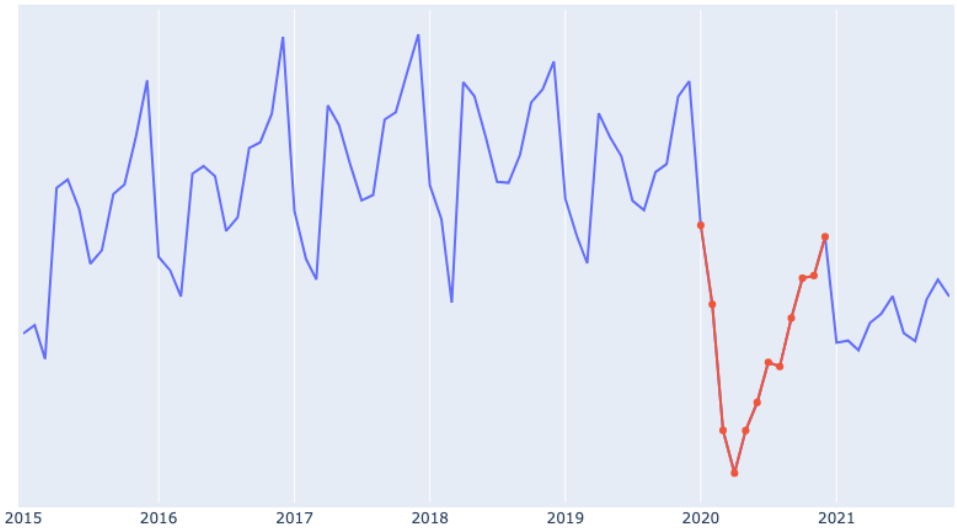

Figure 1.4: Monthly volume sales for a large luxury company

Post-crisis trajectories vary by industry and business. Companies are now

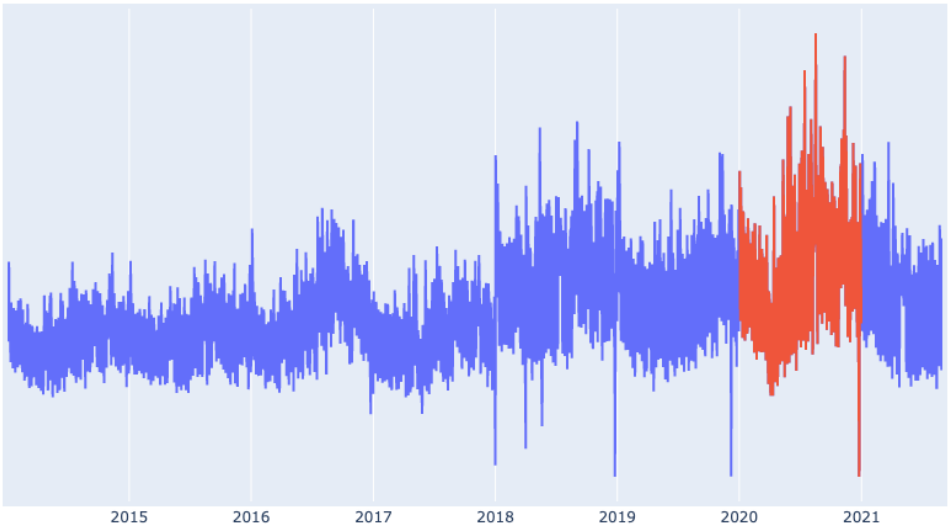

Figure 1.5: Daily volume of calls in a call center for a telecommunication company

entering the next phases to recover and reinvent.

As forecasting relies both on past data and external regressors to predict future, the ongoing crisis also challenges the effective delivery of accurate predictions.

So what are the best practices when using time series forecasting techniques to predict the evolution of key business indicators, like sales or costs?

1.3 Purpose

Data scientists who are experiencing drop of performance for their forecasting models when including COVID-19 period data points should benefit from this degree project by learning how to detect the anomalies and exploring several methods to restore model's predictive power. Forecasting accurately is essential at many levels. The pursuit of accurate forecasting aligns closely with the principles of sustainability. When forecasting accurately, companies are better equipped to optimize their operations and minimize waste, leading to more sustainable outcomes.

One key aspect of accurate forecasting is the ability to anticipate demand

with precision. By understanding customer needs and preferences, companies can avoid overproduction and reduce excess inventory. This, in turn, helps to lower carrying costs and prevents the wasteful disposal of products with limited shelf-life. By aligning production levels with actual demand, companies can significantly reduce their environmental impact, conserving resources and reducing energy consumption.

Accurate forecasting also enables companies to minimize stock-outs, where products are unavailable when customers want to purchase them. By accurately predicting demand, businesses can ensure sufficient inventory levels to meet customer needs, reducing the likelihood of stock-outs. This not only improves customer satisfaction but also eliminates the need for rush orders or expedited shipping, which can contribute to increased carbon emissions and unnecessary resource consumption.

Moreover, accurate forecasting promotes efficient supply chain management. By forecasting demand accurately, companies can optimize their logistics, transportation, and distribution processes. This allows for better route planning, load optimization, and reduced transportation mileage, resulting in lower fuel consumption and greenhouse gas emissions. Sustainable practices such as consolidated shipments and route optimization can be implemented based on accurate forecasts, minimizing the carbon footprint associated with transportation.

Additionally, accurate forecasting contributes to sustainable product development. By understanding future demand trends, companies can invest in the development of sustainable products and services that align with evolving consumer preferences. This proactive approach to product development reduces the risk of producing goods that may become obsolete or have a negative environmental impact. Accurate forecasts can also guide companies in incorporating sustainable materials, manufacturing processes, and supply chain practices into their product design and development, further promoting sustainability throughout the entire product lifecycle.

In summary, accurate forecasting plays a vital role in promoting sustainability. It enables companies to optimize their operations, reduce waste, conserve resources, minimize environmental impact, and develop sustainable products and services. By incorporating accurate forecasting practices into their operations, businesses can contribute to a more sustainable future, aligning their goals with the growing global focus on environmental stewardship.

The topic of accurate forecasting and its alignment with sustainability can be associated with several United Nations Sustainable Development Goals (SDGs).

SDG 9 (Industry, Innovation, and Infrastructure) is relevant as accurate

forecasting contributes to efficient and sustainable industrial practices by optimizing production processes, reducing waste, and improving resource allocation.

SDG 12 (Responsible Consumption and Production) is particularly relevant as accurate forecasting helps companies minimize waste, reduce overproduction, and align production levels with actual demand, promoting sustainable consumption and production patterns.

SDG 13 (Climate Action) is also applicable as accurate forecasting allows companies to optimize supply chains, reduce transportation mileage, and lower carbon emissions, contributing to mitigating climate change.

1.4 Goals

The goal of this project is to provide a guideline for dealing with COVID-19 data points for forecasters. This has been divided into the following two sub-goals:

1. Understand how the COVID-19 crisis has impacted demand on some use cases and provide a clear methodology to analyse and quantify this impact

2. Understand how the COVID-19 crisis has impacted the performance of models trained on data containing the COVID-19 crisis (i.e. during lockdowns) and provide a clear methodology to improve forecasting accuracy despite atypical observations due to COVID-19 crisis in the training set.

1.5 Research Methodology

The first part of this degree project will be dedicated to introducing a bit of background in time series forecasting to understand why abnormalities like COVID-19 crisis can disturb models predictive power.

Then a couple of selected methods will be explored to both qualify the impact of COVID-19 on forecasting models and to restore performance.

Finally, the performance of these methods will be compared on a simple univariate time series representing the sales in clothing in retail in the UK and on a more complex time series also sales data in retail, but with multiple stores and multiple groups of products.

1.6 Delimitations

This degree project provides a methodology to detect and handle forecasting models performance drop due to training on data presenting abnormalities during COVID-19 crisis.

The aim is to increase forecasting models performance when trained on COVID-19 data by making the hypothesis that the future will probably come back to what it was before the crisis. The goal is not to predict future crisis which is out of scope.

The methodologies presented are not exhaustive, nor will work for every use case. Results on different use cases will be presented in the last chapter.

1.7 Structure of the thesis

The thesis is broken down as follows:

1. Chapter 2 presents relevant background information about time series forecasting.

2. Chapter 3 presents what effects were observed following covid-19 by the community, and what are some methods that were used to deal with it.

3. Chapter 4 presents the methodology and methods used to explore how COVID-19 has impacted demand and provides some examples on specific uses cases.

4. Chapter 5 presents how COVID-19 has impacted forecasting models performance, and provides the methodology and methods to deal with atypical observation.

5. Chapter 6 presents the results and some suggestions to improve the methods presented.

6. Chapter 7 presents the conclusions and future work.

Chapter 2

Time series forecasting

Before deep diving into how new observations of COVID-19 pandemic period have impacted the estimated parameters of time series forecasting models, we will start first by explaining what is time series forecasting and why abnormal observations like the COVID-19 pandemic can lead to predicting future values with poor forecast accuracy.

This chapter gives an overview of what a time series is, what are the main forecasting models and methods used nowadays.

2.1 Definitions

In [1] *Forecasting: Principles and Practice*, Rob J Hyndman and George Athanasopoulos define **forecasting** as "predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts". **Time series data** is a quantitative data observed sequentially over time, most of the time at regular time intervals, and therefore **time series forecasting** aims at predicting future values based on present and past data. So how do we characterize a time series?

2.2 Time series decomposition

Time series data is often difficult to interpret if it is not split into several components. Often, forecasters split the original time series into a trend-cycle component T_t, seasonal(s) component(s) S_t and a remainder component R_t (containing anything else in the time series). If we assume an additive

decomposition of the time series, in *Forecasting: Principles and Practice* [1] the data y_t is written as follow

$$y_t = T_t + S_t + R_t$$

The trend and seasonal components are defined in *section 2.3 Time series patterns of Forecasting: Principles and Practice* as follow:

- Trend : "exists when there is a long-term increase or decrease in the data".
- Seasonality : "occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known period."

2.3 Time series forecasting methods

Time series methods use historical data as the basis of estimating future outcomes. They are based on the assumption that past demand history is a good indicator of future demand.

Some examples of simple forecasting methods mainly used as a baseline to compare to other forecasting models (see example in *A Naïve Approach for Comparing a Forecast Model* [2]):

- Mean method (forecasts of all future values are equal to the average of the historical data)
- Naïve method (forecasts of all future values are equal to the last observation).
- Seasonal naïve method (each forecasts of future values are equal to the last observed value from the same season of the year)

2.4 Time series forecasting models

In this section we will describe some models that can be used to forecast time series. Here is a non exhaustive list of time series forecasting models:

- Time series regression models (*Time series regression* [3])
- Exponential smoothing (*A Study on Exponential Smoothing Method for Forecasting* [4])

- Autoregressive moving average or ARMA (*Research on Applications of ARMA in Forecasting of Time Series* [5])
- Autoregressive integrated moving average or ARIMA (*Forecasting of demand using ARIMA model* [6])
- Seasonal ARIMA or SARIMA (*Time Series Forecasting of temperatures using SARIMA An average Example from Nanjing* [7])
- Dynamic regression models (*Dynamic regression models and their applications in survival and reliability analysis* [8])
- Gradient boosted trees (*Multivariate Boosted Trees and Applications to Forecasting and Control* [9])
- Recurrent neural network (*Recurrent Neural Networks for Forecasting Time Series with Multiple Seasonality: A Comparative Study* [10])

2.4.1 Multiple linear regression

The linear regression model assumes that there exists multiple linear relationship between the forecast variable y_t and the predictor variables $\boldsymbol{X_t^n}$

$$y_t = \boldsymbol{X_t^n}\boldsymbol{\beta} + \epsilon_t$$

The vector $\boldsymbol{X_t^n} = (X_t^1, ..., X_t^n)$ is the predictor variables, with n being the number of variables. The vector coefficients $\boldsymbol{\beta}$ are estimated using the least squares methods which aims at minimzing $\sum \epsilon_t^2$

2.4.2 Exponential smoothing

The exponential smoothing methods are based on averaging past values of a time series in a decreasing exponential manner. Simple Exponential Smoothing (SES) assumes no trend or seasonal patterns.

$$y_t = \sum_{k=1} \alpha(1-\alpha)^k y_{t-k}$$

where $0 < \alpha < 1$ is the smoothing parameter.

The weights $\alpha(1-\alpha)^k$ decrease exponentially as observations come from further in the past — the smallest weights are associated with the oldest observations.

The larger α is, the more weight is given to the more recent observations.

There exists other versions of exponential smoothing. Holt Exponential Smoothing [11], Damped Exponential Smoothing [12] and Holt-Winters are extensions of SES for handling the short-term trend, the long-term trend and the seasonal component of time series.

2.4.3 ARIMA models

An Auto Regressive Integrated Moving Average (ARIMA) model is composed of an Auto Regressive (AR) and a Moving-Average (MA) part.

The Auto Regressive (AR) part consists of a linear regression on past values of the time series. If you recall, in a multiple regression model, we forecast the variable of interest using a linear combination of predictors. In an autoregressive model, we forecast the variable of interest using a linear combination of past values of the variable.

The Moving-Average (MA) part is a regression on past forecast errors.

ARMA

ARMA (autoregressive moving average) processes are defined by linear difference equations with constant coefficients. We rarely use ARMA process for real life time series, but they are important to define ARIMA processes.

Let's first introduce the backward shift operator B as it is done in Peter J. Brockwell and Richard A. Davis' *Introduction to Time Series and Forecasting* [13] to simplify the equation of ARMA processes. We define the backward shift operator B as :

$$By_t := y_{t-1}$$

So the lag-1 difference operator $\bigtriangledown$ can be written:

$$\bigtriangledown y_t = y_t - y_{t-1} = (1 - B)y_t$$

y_t is an ARMA(p, q) process if y_t is stationary and if for every t:

$$\phi(B)y_t = \theta(B)z_t$$

where $z_t \sim WN(0, \sigma^2)$ is white noise, a sequence of uncorrelated random variables, each with zero mean and variance σ^2.

where $\phi()$ and $\theta()$ are the p and q-degree polynomials

Simple ARIMA

A generalization of ARMA processes, which incorporates a wide range of non-stationary series, is provided by the ARIMA processes, i.e., processes that reduce to ARMA processes when differenced finitely many times.

ARIMA models can therefore handle non-stationarity by differencing the time series. The differencing operation consists basically in computing the difference between a times series at time t and the same times series shifted by a timestep in the past. We note d the number of times that the raw observations are differenced, also known as the degree of differencing.

If d is a non-negative integer, then y_t is an ARIMA(p,d,q) process if :
$w_t := (1 - B)^d y_t$ is a causal ARMA(p, q) process.

This definition means that y_t satisfies a difference equation of the form:

$$\phi(B)(1 - B)^d y_t = \theta(B) z_t$$

and $z_t \sim WN(0, \sigma^2)$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q, respectively, and $\phi(z) \neq 0$ for $|z| \leq 1$.

In other words, p is the order of the AR terms, q the order of the MA terms and d the differencing degree.

Seasonal ARIMA (SARIMA)

ARIMA is not meant to handle seasonalities. Instead, we use Seasonal ARIMA which uses differencing again but this time at lag s to eliminate the seasonal component of period s. D is the number of times we difference the raw observations to remove the seasonal part.

If d and D are non-negative integers, then y_t is a seasonal ARIMA(p, d, q) × (P, D, Q)s process with period s if the differenced series $w_t = (1 - B)^d (1 - B^s)^D y_t$ is a causal ARMA process defined by:

$$\phi(B)\Phi(B^s) y_t = \theta(B)(B^s) z_t$$

and $z_t \sim WN(0, \sigma^2)$

(p,d,q) are the parameters of the classical ARIMA part, whereas (P,D,Q) are the parameters of the seasonal part.

Seasonal with exogenous regressors (SARIMAX)

Time series are sometimes influenced by external variables. For example, if our task is to predict the number of person who will cycle tomorrow, it is sure important to know the historical numbers to make an accurate prediction, but it might also be very helpful to know the weather. If it is sunny tomorrow, there will likely be more cyclists then if it rains.

The above models are simple time series forecasting methods that use only information on the variable to be forecast, and make no attempt to discover the factors that affect its behaviour. Therefore they will extrapolate trend and seasonal patterns, but they ignore all other information such as marketing initiatives, competitor activity, changes in economic conditions, and so on.

Seasonal Auto Regressive Integrated Moving Average with eXogenous regressors (SARIMAX) is a "mixed model" that uses both past values of the variable to predict, but also a linear regression on external variables.

2.4.4 Gradient boosting trees

Gradient boosting trees (GBT) are a category of decision trees ensemble. GBT uses ensemble learning, which means it creates multiple learners and aggregates their result them to make a prediction. More specifically, it uses boosting which involves building simpler weak learners sequentially, where each model tries to predict the error (residuals) left over by the previous model. Each weak learner is a decision tree with usually 8 to 32 leaves.

To predict a set of observations $(y_i)_{i=1}^n$, GBT aggregates a number K of regression trees f_t with weights ρ_t from a set of features $(x_i)_{i=1}^n$.

$$y_i = F(x_i) = \sum_{t=1}^{K} \rho_t f_t(x_i)$$

The value $f_t(x_i)$ depends in which leaf the sample x_i is classified in the tree f_t, and the score associated to that specific leaf of that tree $w_{f_t,leaf}$. GBT trained to minimize the objective function:

$$L = \sum_{i=1}^{n} l(y_i, F(x_i))$$

where l is a loss function (typically Mean Squarred Error).

A drawback of GBT algorithms is that they are prone to overfitting, and thus show poor performance on the test dataset.

XGBoost

XGBoost is one of the most widely decision trees used today because of its good performance (see example of application to time series forecasting in *Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah* [14]). XGBoost is similar to GBT, with some extra feature:

- Pruning to improve generalization
- Extra regularization function in the objective function also to improve generalization

So this time, the objective function can be written:

$$L = \sum_{i=1}^{n} l(y_i, F(x_i)) + \sum_{k=1}^{K} \Omega_k(f_t)$$

where Ω is a regularization function (for example L1 or L2 regularization) to penalize the complexity of the trees and to prevent overfitting issues.

LightGBM

LightGBM is another type of boosting algorithm, which was designed to improve the scalability and reduce computation times. It has shown to be faster and sometimes more accurate than XGBoost (see example of application to time series forecasting in *Supply chain sales forecasting based on lightGBM and LSTM combination model* [15]).

What makes LightGBM different is that it uses a unique technique called Gradient-based One-Side Sampling (GOSS) to filter out the data instances to find a split value. This is different than XGBoost, which uses pre-sorted and histogram-based algorithms to find the best split.

Catboost

CatBoost is a boosting algorithm based on Gradient Descent that has low latency requirements, which translates to it being around eight times faster than XGBoost (see example of application to time series forecasting in *A Crypto Market Forecasting Method Based on Catboost Model and Bigdata* [16]). The main differences:

- CatBoost implements symmetric trees, which help in decreasing prediction time, and it also has a shallower tree-depth by default (six)

- CatBoost leverages random permutations similar to the way XGBoost has a randomization parameter
- Unlike XGBoost however, CatBoost handles categorical features more elegantly, using concepts like ordered boosting and response coding

2.5 Forecasting accuracy

The forecast error (also known as a residual) is the difference between the actual value and the forecast value for the corresponding period:

$$e_t = y_t - \hat{y}_t$$

There exists different types of forecast accuracy metrics, some are scale-dependent errors, like the Mean Absolute Error (MAE), the Mean Squarred Error (MSE) or the Root Mean Squarred Error (RMSE), defined as follow:

$$MAE = (e_t)$$
$$MSE = \mathbb{E}(e_t^2)$$
$$RMSE = \sqrt{\mathbb{E}(e_t^2)}$$

MSE is serves usually as a loss function for regression. MAE and RMSE are serve usually as validation metrics. Both give insights of how good the model performs by looking at how much the predictions deviate from the ground truth. However, the main difference between MAE and RMSE is that RMSE penalizes more large residuals than small ones because of the squarre. However, these metrics cannot be used to compare forecast accuracy across series with different units.

For that, we can compute percentage errors like the Mean Absolute Percentage Error (MAPE) that are unit free, defined as follow :

$$MAPE = \mathbb{E}(100 * |\frac{e_t}{y_t}|)$$

We introduce a custom metric FA that we will refer as forecast accuracy, defined as follow:

$$FA = 1 - \frac{\sum |e_t|}{\sum y_t}$$

The forecast accuracy's value ranges from 0 to 1, the closest it is to 1 the better the forecast.

We will also measure the bias:

$$bias = \frac{\sum e_t}{\sum y_t}$$

A large negative bias means that the model tends to under forecast and a large positive bias means that the model tends to over forcast. The closest it is to 0 the better the forecast.

Chapter 3

Dealing with COVID-19 in forecasting models by the community

In the previous chapter, we introduced what are the main forecasting models used today, and how they use historical data to provide future predictions.

As introduced in the first chapter, COVID-19 pandemic has altered the demand data patterns. The problem of using historical data and not taking into account the COVID crisis is that the forecasts will likely still rely on the assumption that many of the underlying behavioral and societal trends are the same as before. This will therefore make data scientist struggle to give accurate forecasts for the future.

In order to generate accurate forecast, one idea would be to accommodate the underlying dynamics of COVID-19 in the statistical modeling process by having a strong hypothesis on the structural impact of COVID-19 on the time series. Another would be to change the COVID-19 data points distribution so it matches what is was before the crisis. Both sides have been explored through several methods that will be discussed in the next section.

3.1 Context

Pandemic presented a major disruption in the demand of a wide variety of products and services. For some sectors there was a short term disruption in demand, whereas for others there was a long term impact in the form of a sustained surge or drop in demand.

During disruptive times, forecasting models which are build using historical data, are challenged by structural changes.

In other words, if there is a significant difference between new and historical values, typical machine learning models are often rendered useless.

3.2 Previous works

3.2.1 Identifying different scenario

After the COVID-19 crisis, three major scenarios can happen regarding the evolution of business indicators [17]:

- Minor impact scenario

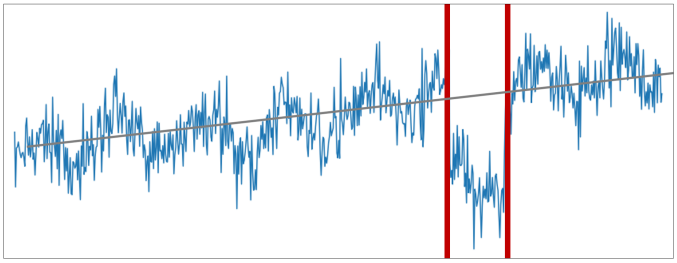

Figure 3.1: Minor impact

In this scenario, the impact is localised. The level, trend and seasonality of the time series is conserved after the COVID-19 crisis. For that, the data points during COVID-19 crisis are irrelevant for future forecast. Predictions are made only based on data points previous to the crisis, so the best practice is to filter out the data points impacted by the crisis.

- Lasting impact scenario

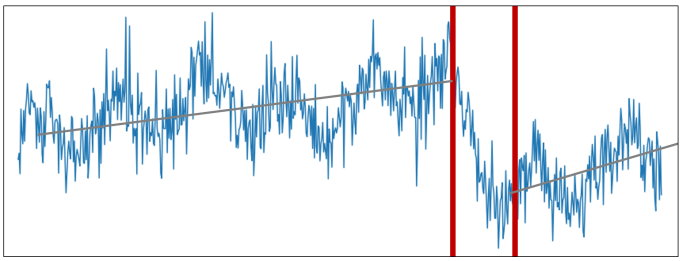

Figure 3.2: Lasting impact

In this case, after a localised major impact during the COVID-19 crisis, the activity resumes but the level and trend of the time series is different

then what it was before the crisis. The seasonal component is however preserved. Here the goal is to evaluate the trend evolution with business experts or by using newest data points post COVID-19 crisis, and keep the predictions of the seasonal component pre COVID-19 crisis.

- Major disruption scenario

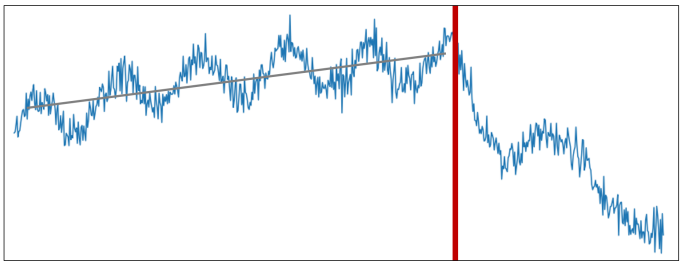

Figure 3.3: Major disruption

In the scenario the current and future behaviors do not reproduce any known, past patterns. Trend and seasonal variations are now totally different. The best practice there consists in discarding most of the known history and consider only the more recent data points for training the time series forecasting model.

In this project, we will therefore study uses cases with a lasting impact scenario or a minor impact scenario, since for the last scenario the only way to get accurate forecasts post COVID-19 is to get rid of data previous to COVID-19 and get more recent data points.

3.2.2 Modeling the underlying dynamics of COVID-19

During the pandemic, most of the literature focused on models to predict the progression of the disease. However, very soon after global lockdowns in 2020, many companies expressed the need for models to predict the impact of COVID-19 disease spread and government regulations on consumer behavior.

As a result, more and more scientific papers have focused on the subject.

For example, PwC studied a quantitative behavior model of fear of COVID-19, impact of government interventions on consumer behavior, and impact of consumer behavior on consumer choice and hence demand for goods in the paper [18] by using panel regression to understand the drivers of demand during the pandemic and Bayesian inference to simplify the regulation landscape that can help build scenarios for resilient demand planning.

On the other hand, Srihari Jaganathan [19] uses Time series models ARIMA transfer function and exponential smoothing to model Taiwan tourist arrivals from Japan. To deal with the impact of COVID-19 on the time series, he uses a method called interrupted time series modeling. The impact of the SARS virus pandemics on tourism is modeled with a compound effect which is a combination of a gradual decline, then a plateauing effect and followed by an exponential increase

Finally, in [20] the authors focused on developing a model that combines personal mobility with motor gasoline demand and uses a neural network to correlate personal mobility with the evolution of the COVID-19 pandemic, government policies and demographic information.

3.2.3 Data processing to correct COVID-19 outliers

Another approach adopted by some data scientists consists in modifying historical outlier data due to COVID-19 instead of trying to find an underlying model to represent the impact of COVID-19 on the target value. This simple approach is sometimes sufficient to train a model that has a good forecast accuracy on post COVID-19 target value.

Several methods to modify historical data were compared [21]:

- Imputation
- Optimal transport
- Time series decomposition

All of these methods consist in modifying an abnormal time range of observations. Imputation consists of erasing the history of the abnormal time ranges (lockdowns for example) to reconstruct a history based on the rest of the available data (previous year data for example).

Optimal transport consists in modifying the part of the abnormal history data called source context (lockdown for example) to transport it to a target context (pre COVID-19 for example). The transformation is based on a principle of optimization under constraint, allowing the transition from one distribution to another by the shortest path [22]. We will go more deeply into the theory of optimal transport in the next section.

Time series decomposition consists in acting separately on each component of the time series during the abnormal time range (source context) for it to match the target context [21]. The operations carried out are as follows:

- The seasonal pattern of the target context is preserved and applied to the source context.
- The trend is extrapolated from the target context either by linear regression or by applying an optimal transport method.
- The source context residuals are normalized, then de-normalized via the inverse normalization operation performed on the target context. This approach makes it possible to adjust the level of the residuals while preserving their nature and behavior.

3.2.4 Optimal transport

Optimal Transport (OT) aims at finding the most efficient way to move mass between distributions [23].

The mathematical problem was introduced by Gaspard Monge in 1781. It was formulated to find a solution to the following problem: How to move dirt from one place (déblais) to another (remblais) while minimizing the effort? The goal is to find a mapping T between the two distributions of mass (transport) while optimizing with respect to a displacement cost $c(x, y)$ (optimal).

Mathematical formulation The optimization problem can be expressed for two probability measures (distributions) μ_s and μ_t and a cost function $c : \Omega_s \times \Omega_t \rightarrow \mathbb{R}$ as:

$$inf_{T \neq \mu_s = \mu_t} \int_{\Omega_s} c(x, T(x)) \mu_s(x) dx$$

where $c(.,.)$ is the ground cost and the constraint $T \neq \mu_s = \mu_t$ ensures that μ_s is completely transported to μ_t.

When working on discrete distributions, the Monge problem can be expressed as a Linear Problem (that always has a solution) with the Kantorovitch formulation (1942) that seeks for a probabilistic coupling $\gamma \in P(\Omega_s \times \Omega_t)$ between Ω_s and Ω_t:

$$\gamma_0 = argmin_\gamma \int_{\Omega_s \times \Omega_t} c(x, y) \gamma(x, y) dx dy$$

s.t. $\gamma \in P = \left\{ \gamma \geq 0, \int_{\Omega_t} \gamma(x, y) dy = \mu_s, \int_{\Omega_s} \gamma(x, y) dx = \mu_t \right\}$

Where γ is a joint probability measure with marginals μ_s and μ_t.

When μ_s and μ_t are only accessible through discrete samples, the corresponding empirical distributions can be written as:

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta_{x_i^s}, \mu_t = \sum_{i=1}^{n_t} p_i^t \delta_{x_i^t}$$

Where δ_{x_i} is the Dirac function at location $x_i \in \mathbb{R}$, p_i^s and p_i^t are probability masses associated to the i-th sample and belong to the probability simplex, $i.e.$ $\sum_{i=1}^{n_s} p_i^s = \sum_{i=1}^{n_t} p_i^t 1$.

In the discrete case of the Kantorovich formulation of the optimal transport problem, we denote β the set of probabilistic couplings between the two empirical distributions defined as:

$$\beta = \left\{\gamma \in (\mathbb{R}^+)^{n_s \times n_t} | \gamma 1_{nt} = \mu_s, \gamma^T 1_{ns} = \mu_s\right\}$$

where 1_d is a d-dimensional vector of ones.

The Kantorovitch formulation of the optimal transport reads:

$$\gamma_0 = argmin_{\gamma \in \beta} \langle \gamma, C \rangle_F$$

where $\langle ., . \rangle_F$ is the Frobenius dot product and $C \geq 0$ is the cost function matrix, whose term $C(i,j) = c(x_i^s, x_j^t)$ denotes the cost to move a probability mass from x_i^s to x_j^t.

The linear program computational complexity is $O((n_s + n_t) n_s n_t \log(n_s + n_t))$. For that reason, Cuturi proposed to regularize the expression of the optimal transport problem by the entropy of the probabilistic coupling. The regularized version of the optimal transport problem is therefore:

$$\gamma_0 = argmin_{\gamma \in \beta} \langle \gamma, C \rangle_F + \lambda \Omega_s(\gamma)$$

where $\Omega_s(\gamma) = \sum_{i,j} \gamma(i,j) \log_{\gamma(i,j)}$ computes the negentropy of γ.

OT for domain adaptation From the optimization problem, one way to use the OT solution to find the optimal mapping (Monge mapping, OT matrix), which aims at finding correspondences between distributions. This way we can find a mapping that can then be used to transfer knowledge between distributions. We will use the optimal mapping in the case of domain adaptation as described in this article [22].

This is a method to change as less as possible each datapoints from the source distribution for it to have the same distribution as the target distribution.

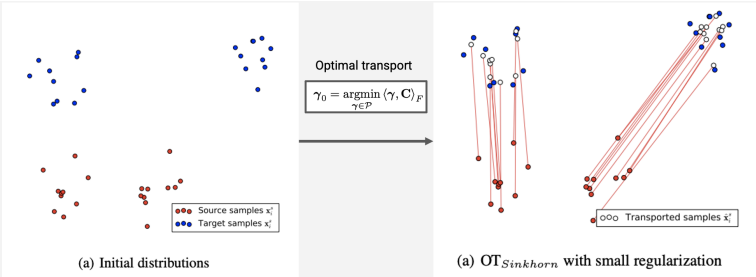

Figure 3.4: Optimal transport for domain adaptation

In the context of domain adaptation, once the probabilistic coupling γ_0 has been computed, source samples have to be transported in the target domain.

The novel distribution $\hat{\mu}$ is a distribution with the same support of μ_t :

$$\hat{\mu} = \sum_j \hat{p_j}^t \delta_{x_i^s}$$

with $\hat{p_j}^t = \sum_i \gamma_0(i,j)$.

The weights $\hat{p_j}^t$ can be seen as the sum of probability mass coming from all samples $\{x_i^s\}$ that is transferred to sample $\{x_j^t\}$.

We can exploit this information to compute a transformation of the source samples. This transformation can be conveniently expressed with respect to the target samples as the following barycentric mapping:

$$\hat{x_i^s} = argmin_{x \in \mathbb{R}^d} \sum_j \gamma_0(i,j) c(x, x_j^t)$$

where x_i^s is a given source sample and $\hat{x_i^s}$ is its corresponding image.

Practical applications Although the initial motivations of Monge and Kantorovitch were respectively military and economic, the optimal transport finds numerous practical applications.

One of the applications is image processing, for example for image color adaptation which consists in transforming the color palette of an image to transport it to the color palette of another image by considering the cost matrix as the euclidean distance between the pixels of the two images.

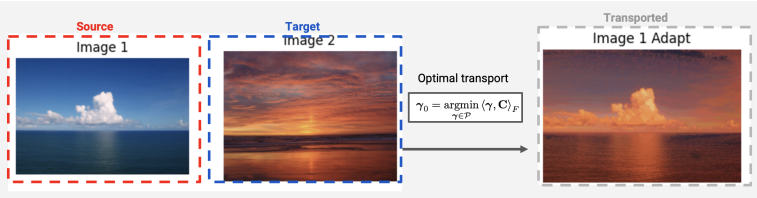

Figure 3.5: Optimal transport for image color adaptation

Chapter 4

Identifying and quantifying the impact of COVID-19 on forecasting models

The first step towards getting an increased forecast accuracy on post COVID-19 observations, is to identify and quantity the impact of COVID-19 on forecasting models.

The effect that had COVID-19 on demand for example is not straightforward. There seems not to be a linear relationship between the number of COVID-19 cases and retail sales for example. Sales depend also on the government measures which are more difficult to quantify.

That is why it is crucial to ask and answer the right questions during the exploratory data analysis (EDA) phase so we can after use the appropriate method to deal with COVID-19 impact on the time series.

4.1 Time series data

To illustrate this section, I chose to work on retail data. The univariate time series represents the sales in clothing in retail in terms of value (amount spent) each month from January 1988 to December 2021 in the UK. The value estimates reflect the total turnover that businesses have collected over a standard period.

Click here to see the source of the data.

4.2 Initial observations

We can see that there is an increasing trend over the years, and a strong seasonality (minimum in February, maximum in December). Also we can observe the strong drop of sales from March 2020 until April 2021.

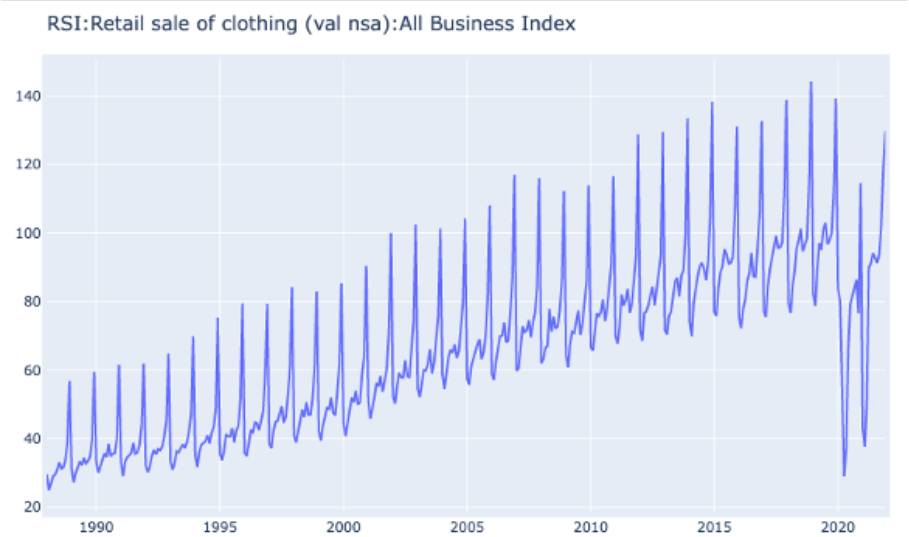

Figure 4.1: Retail sale of clothing

To be convinced that COVID-19 should be considered into modeling, we will compare our predictive performance on a train/test set previous COVID-19 and on a train/test set that include the COVID-19 impact period.

For that, we will model our time series with a Facebook Prophet model [24] which is based on an additive model where non linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It provides completely automated forecasts.

To keep it simple, we will model our time series with a linear trend, a yearly seasonality and a holiday effect specific for Great Britain (GB).

Prophet package allows us to decompose the different components of the model : the trend, the seasonality and the holiday effects.

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

where :

- g(t) is the trend function which models non-periodic changes in the value of the time series

- s(t) represents periodic changes (e.g., weekly and yearly seasonality)
- h(t) represents the effects of holidays which occur on potentially irregular schedules over one or more days
- The error term ϵ_t represents any idiosyncratic changes which are not accommodated by the model, and is normally distributed.

4.2.1 Before COVID-19

We split the data into a train set and a test set. The train set starts in January 1988 and ends in June 2019, whereas the test set starts in July 2019 and ends in December 2019 just before COVID-19 pandemic (6 month of test data).

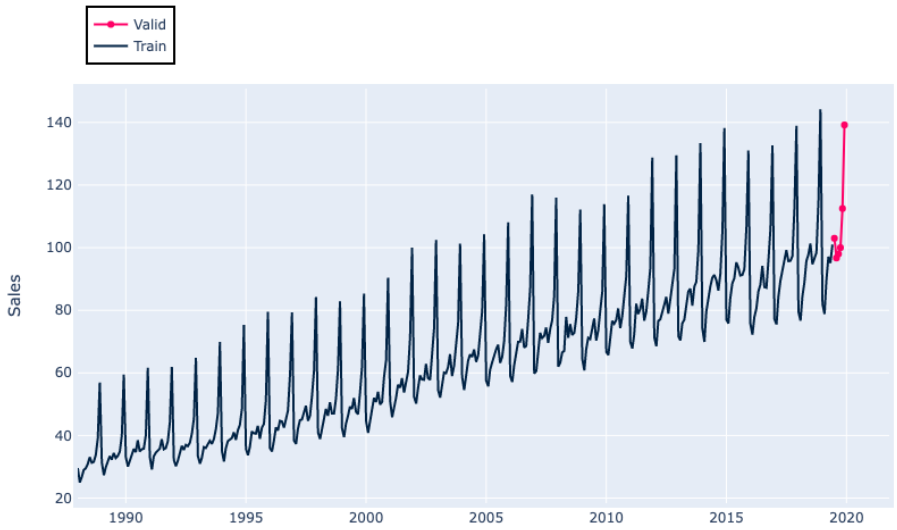

Figure 4.2: Train/Test split before COVID-19

We then fit a prophet model on the train set, and measure the forecast accuracy and the bias of the model's predictions (as defined in the section 2.4) on the test set.

We get the following results:

If we get a look at the predictions, we can see that the model fits very well on the training set and generalizes well on the test set, which is reflected by the forecast accuracy very close to 1 and the bias very close to 0.

We can see that the increasing linear trend is well captured, as well as the yearly seasonality with a strong increase in sales in December.

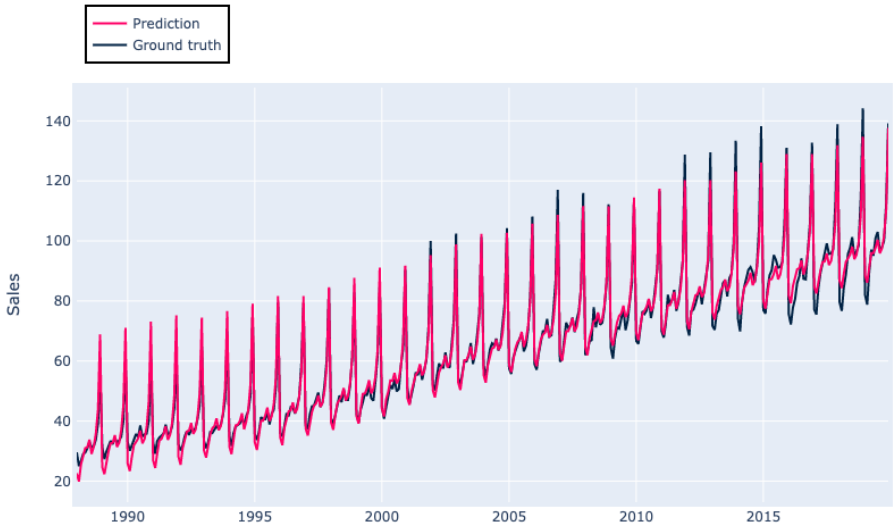

Figure 4.3: Predictions before COVID-19

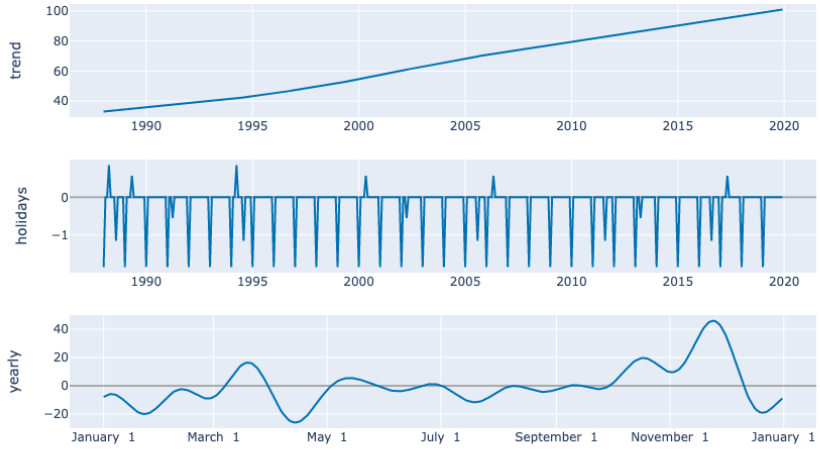

Figure 4.4: Time series components before COVID-19

Table 4.1: Before COVID-19 model performance

Train	Test	Forecast Accuracy	Bias
Jan 1988-June 2019	July 2019-Dec 2019	0.986	-0.00917

4.2.2 After COVID-19

Including COVID-19 period data

Now if we use all the data available (which includes the COVID-19 period), the train set starts in January 1988 and ends in June 2021, whereas the test set starts in July 2021 and ends in December 2021 including major COVID-19 lockdowns (6 month of test data).

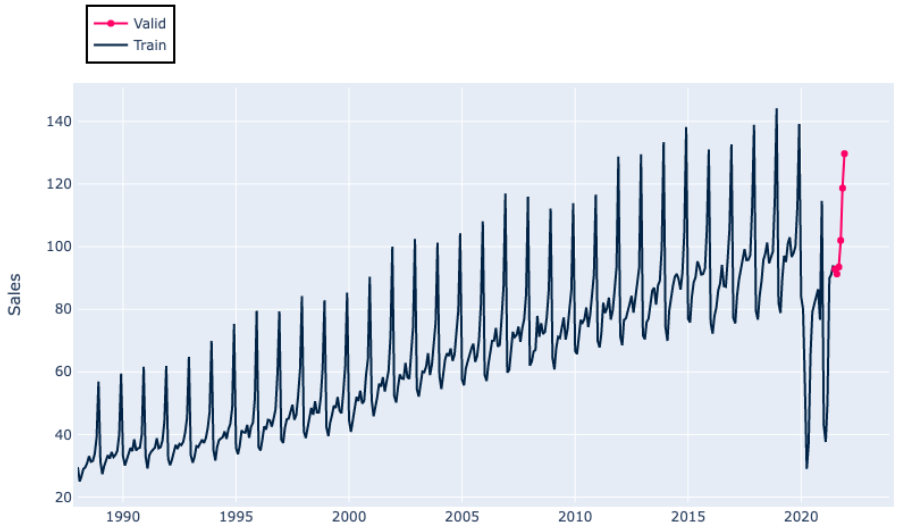

Figure 4.5: Train/Test split after COVID-19

We then fit a prophet model on the train set, and measure the forecast accuracy and the bias of the model's predictions (as defined in the section 2.4) on the test set.

We get the following results:

This time, with the abnormal COVID-19 period included in the training set, the model under forecasts on the test set (we underestimate future demand predictions compared to the true values). Therefore, the forecast accuracy drops to 0.903 and the bias moves towards negative values almost ten times

Table 4.2: After COVID-19 model performance

Train	Test	Forecast Accuracy	Bias
Jan 1988-June 2021	July 2021-Dec 2021	0.903	-0.0887

larger then what we had before COVID-19.

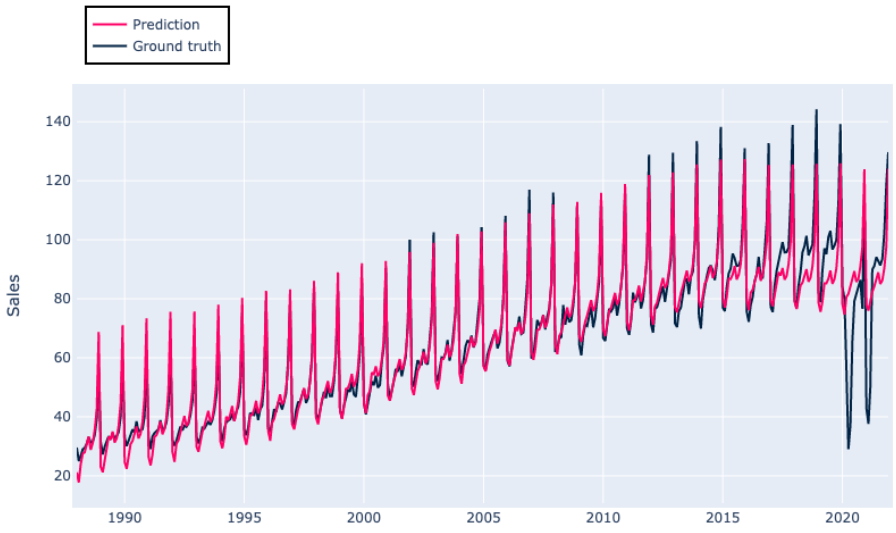

Figure 4.6: Predictions after COVID-19

We can see that the trend isn't only composed of an increasing linear trend anymore, but it seems to slowly decrease from 2015. Also the yearly seasonality seems to have changed, with March being also a strong sales month. We can clearly see that Prophet model is struggling to find a way to model the COVID-19 effect by including it in the trend and the seasonality.

Removing COVID-19 period data

If our goal is to make post COVID-19 forecast more accurate, and we are in a case of a minor impact scenario where the COVID-19 period is localised on a specific time period, then why not removing the COVID-19 period from the training set? We would be taking the hypothesis that the COVID-19 impacted the time series structure only on a localised time period, and that will not affect future values.

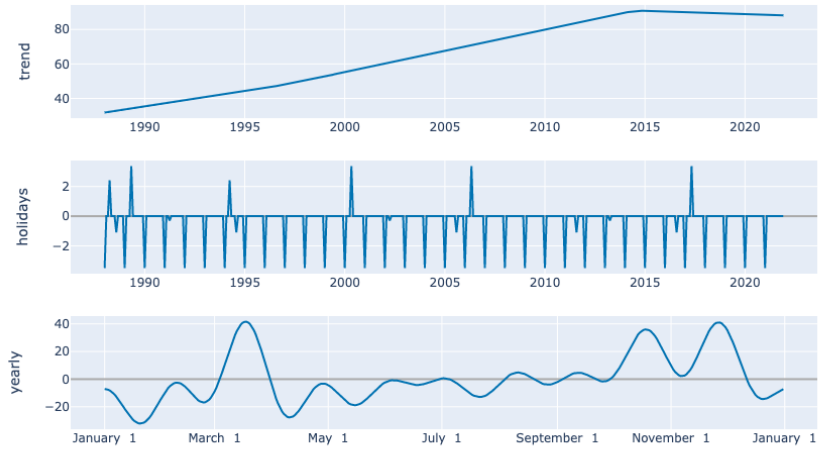

Figure 4.7: Time series components after COVID-19

The train set starts in January 1988 and ends in February 2020 (excluding March 2020 till June 2021), whereas the test set starts in July 2021 and ends in December 2021.

We then fit a prophet model on the train set, and measure the forecast accuracy and the bias of the model's predictions (as defined in the section 2.4) on the test set.

We get the following results:

Table 4.3: After COVID-19 model performance (excluding COVID-19 period)

Train	**Test**	**Forecast Accuracy**	**Bias**
Jan 1988-Feb 2020	July 2021-Dec 2021	0.935	0.0507

We have increased the forecast accuracy compared to the previous model which included the COVID-19 period in the training set, but this time the model over forecasts on the test set as the bias moves towards positive values.

We have however corrected the trend as it is now only linear increasing. Also the yearly seasonality is almost back to what it was before COVID-19.

So we can see that this simple approach of removing the COVID-19 period is better for having a better forecast accuracy and bias on future predictions in a post COVID-19 world. But we can also notice that the model is not completely

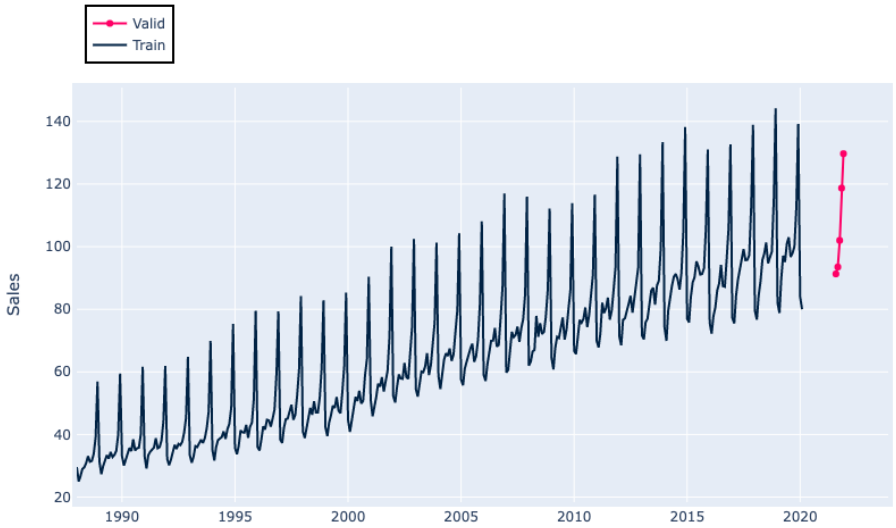

Figure 4.8: Train/Test split after COVID-19 (excluding COVID-19 period)

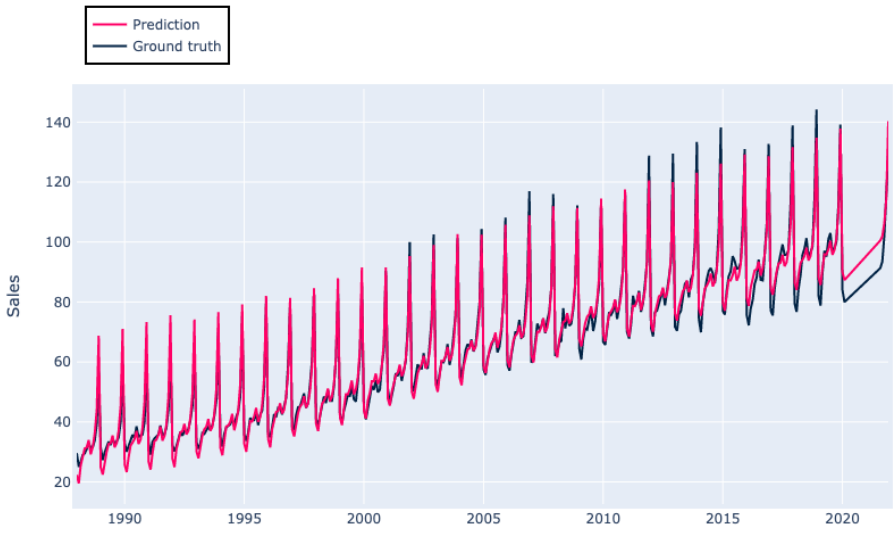

Figure 4.9: Predictions after COVID-19 (excluding COVID-19 period)

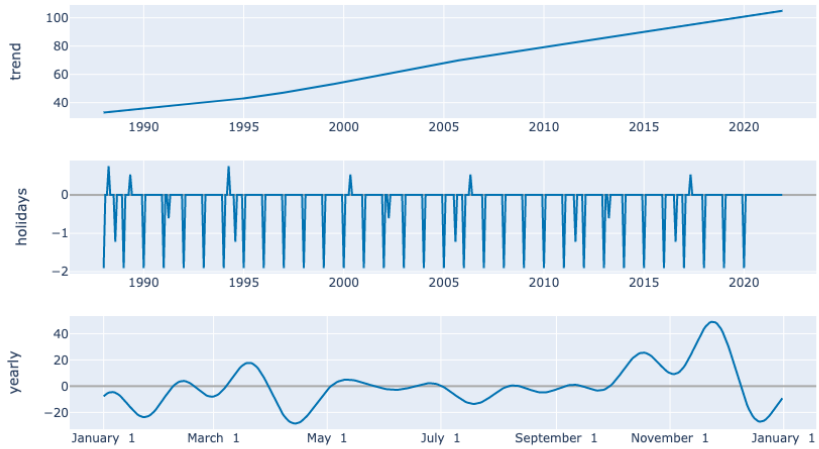

Figure 4.10: Time series components after COVID-19 (excluding COVID-19 period)

accurate as we are losing some information by removing the COVID-19 period. The trend has changed, is slightly less increasing then before, but not decreasing neither.

What is a better approach for this problem?

4.3 Exploratory Data Analysis (EDA)

Before exploring methods to correct the COVID-19 effect on forecasts, we must establish a clear methodology to understand the effect on forecasts, identify the data points considered as anomalies that require a specific data processing method before being used as training data.

For that, we came up with an issue tree for demand predictions use cases that can guide data scientists during the data processing steps. This issue tree was designed for a specific use case (a retail company selling different products across different stores) but can be adapted to different use cases (PRODUCT and STORE correspond to 2 different hierarchies).

Each question gives a hint to how and what data processing method to use. For example, by identifying the time range that gives abnormal total sales volume, we can identify the data points that will need data processing. By identifying the stores impacted by COVID-19, we can decide if we apply the

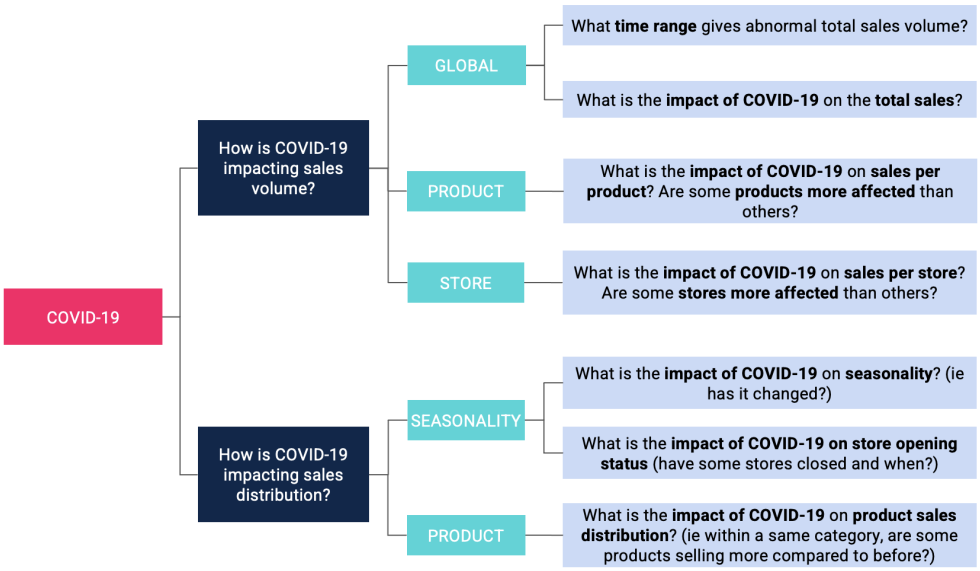

Figure 4.11: EDA issue tree

same data processing to all stores or to some specific stores.

In the following subsections, we will explore each branch of the issue tree and try to provide an answer to the question.

4.3.1 How is COVID-19 impacting total sales volume?

In this section, we will evaluate COVID-19 impact on the times series level.

What time range gives abnormal total sales volume?

We can identify the time range of the COVID-19 impacted data by performing some transformations on the time series. These transformations are different in the case where the time series has a clear trend.

When there is no clear trend For this question, we could observe directly the total target variable (in the case of the retail company, the target variable is the sales quantity per day per store per product) against time and identify the anomalies. But this method does not work when it is hard to identify a pattern. Thus we came up with another method.

To identify the anomalies, we first define a baseline. A baseline is a set of data points that are considered as 'normal' (representative of the 'normal'

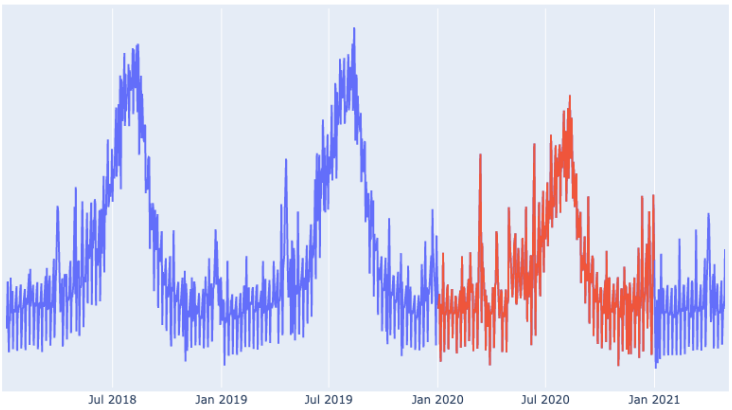

Figure 4.12: Target variable (sales quantity per day per product per store) against time for a retail company. The red line represents year 2020.

behaviour of the target variable against time). For the retail company, the baseline goes from January 2018 till December 2019. Once the baseline has been identified, we can then visualize the growth rate by month. The function aggregates first the target variable value per month, and then calculates a growth rate defined as the ratio between the current value and the baseline value (average target variable value during the baseline period for that specific month). We then get the following graph where the red line represents the baseline, and the green lines represent the baseline plus and minus three times the standard deviation.

If the time series is stationary, then the noise left should follow a normal distribution. Therefore, we can admit that if the growth rate value at a certain time is greater than $baseline + 3 * std$ or lower than $baseline - 3 * std$ than it is an abnormality. We can clearly see that the impact of COVID-19 is strongest from April 2020 till December 2020.

We follow the same methodology for two other different use cases : a luxury company wanting to predict sales of new innovation products and a telecommunication company wanting to predict the number of calls in its call centers.For the luxury company, the major impact time period can be identified as February 2020 - November 2021. The COVID-19 effect on sales is a lasting impact, as 2021 seems not quite back to normal.

Whereas for the telecommunication company, the time range of abnormal observations is not very clear.

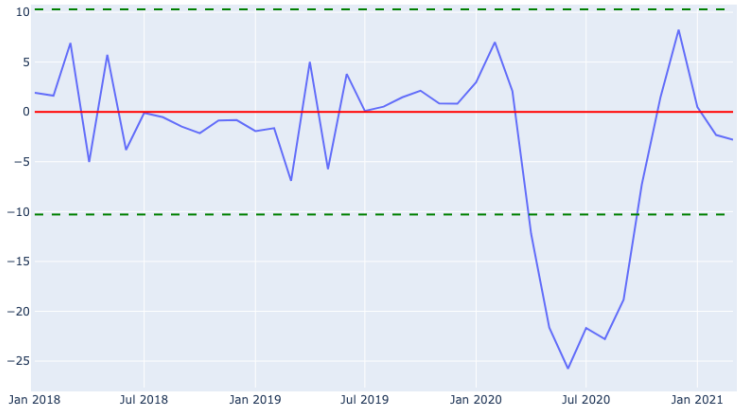

Figure 4.13: Growth rate by month (between current value and baseline value) against time for the retail company

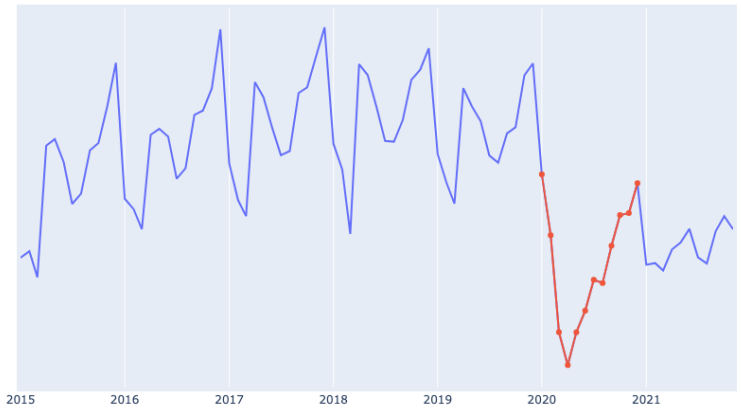

Figure 4.14: Target variable (sales quantity per day per product per channel) against time for a luxury company. The red line represents year 2020.

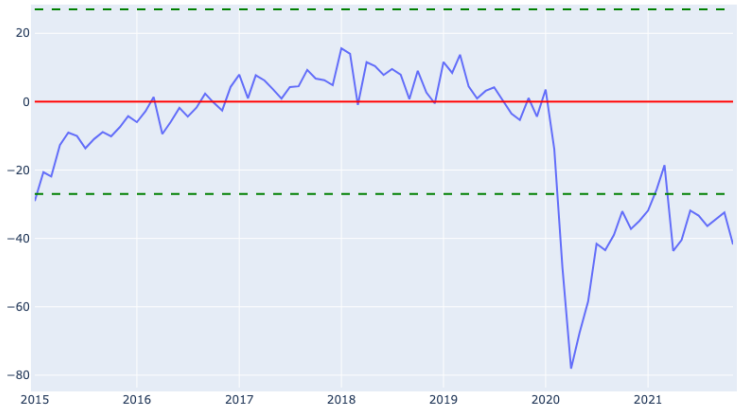

Figure 4.15: Growth rate by month (between current value and baseline value) against time for the luxury company

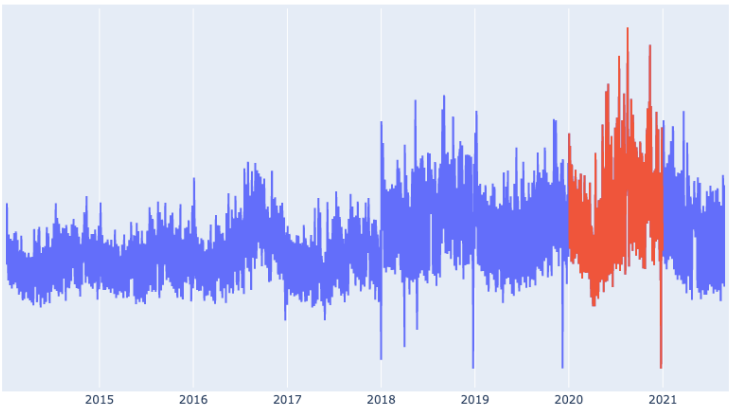

Figure 4.16: Target variable (number of calls per day per call center) against time for a telecommunication company. The red line represents year 2020.

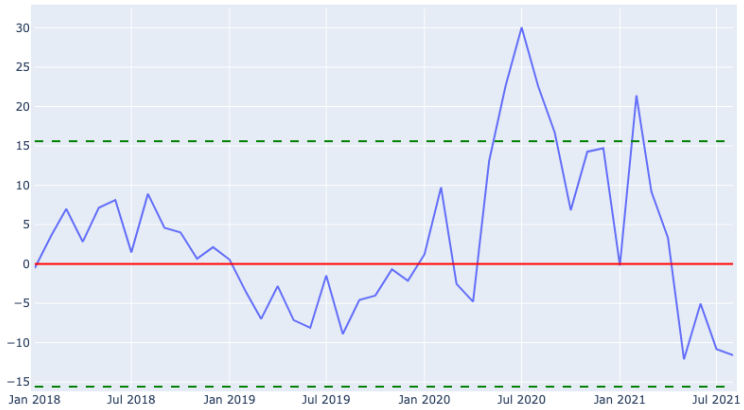

Figure 4.17: Growth rate by month (between current value and baseline value) against time for the telecommunication company

When there is a clear trend When there is a clear trend in the time series, like it is the case for the retail sale of clothing in the UK time series, we must first remove the trend to better understand the impact of COVID-19 on the target variable. The trend of this time series is clearly linear.

Removing the trend - To remove the trend we can chose between two methods :

- Differentiate : Perform a first order differentiation (actual value minus previous value).
- Fitting a linear trend

For the second method:

- We first fit a linear model (sklearn LinearRegression) on the whole time series.
- We then remove the predicted linear trend from the target variable, which renders a detrended time series.

On the detrended time series, we can already see the 2020 COVID-19 crisis with an abnormality streching from April 2020 till July 2021.

Removing the seasonality - We make the last observation even more obvious by removing the seasonality in the data. As we saw in the previous visualization the same pattern seems to repeat itself every year. That's due to seasonality (sales are stronger in December during the Christmas period). To

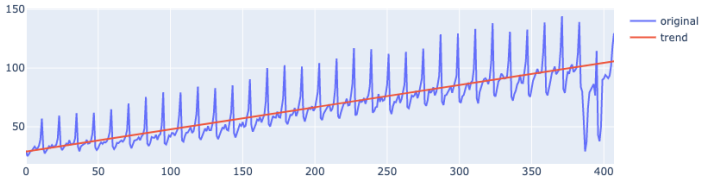

Figure 4.18: Fit a linear model to the sale of clothing time series

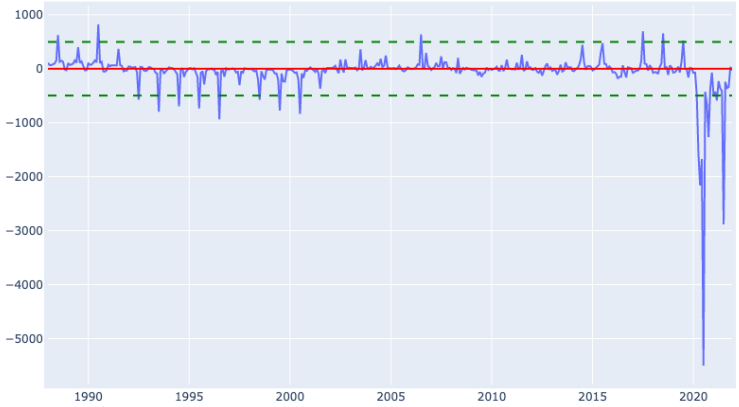

Figure 4.19: Detrended time series

remove it, we will compute an average of sales per month on the detrended sales before 2020 and then remove that value from the sales value. Therefore, we will have a detrended and seasonally adjusted time series.

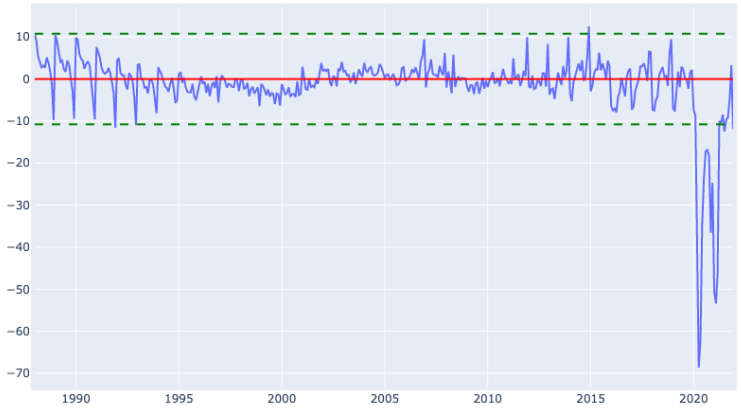

Figure 4.20: Detrended and seasonally adjusted time series

We have now nearly a stationary time series. We can observe however that until 1995 December had a lower seasonality than later (Christmas sales). If the time series is stationary, then the noise left should follow a normal distribution. Therefore, we can admit that if the target value at a certain time is greater than $mean + 3 * std$ or lower than $mean - 3 * std$ than it is an abnormality. We can clearly see that the impact of COVID-19 is strongest from March 2020 till April 2021.

What is the impact of COVID-19 on the total sales?

- Retail company: a decrease of sales volume during summer, up to -25% per month
- Luxury company : a decrease of sales volume during since the first lockdown, up to -78% per month.
- Telecommunication company: an sudden increase of calls in February 2020 (Lockdown announce), followed by a sudden decrease in March and April (Lockdown), and a sudden increase of calls from May till end of Summer (lift of lockdown). Not returned to normal.

What is the impact of COVID-19 on sales per product? Are some products more affected than others? We have identified the global impact of COVID-19 on sales during the time range identified as abnormal. This time, we try to identify the categories of products that were most affected by COVID-19. To do so, we proceed the same way as before, but this time we define a baseline per category of product per year. We then can visualize the growth rate in 2020 compared to baseline (before 2020) for different categories of products. The purpose of this analysis is to understand if all categories / channels are identically impacted by COVID-19 to identify the scope of the data processing transformation.

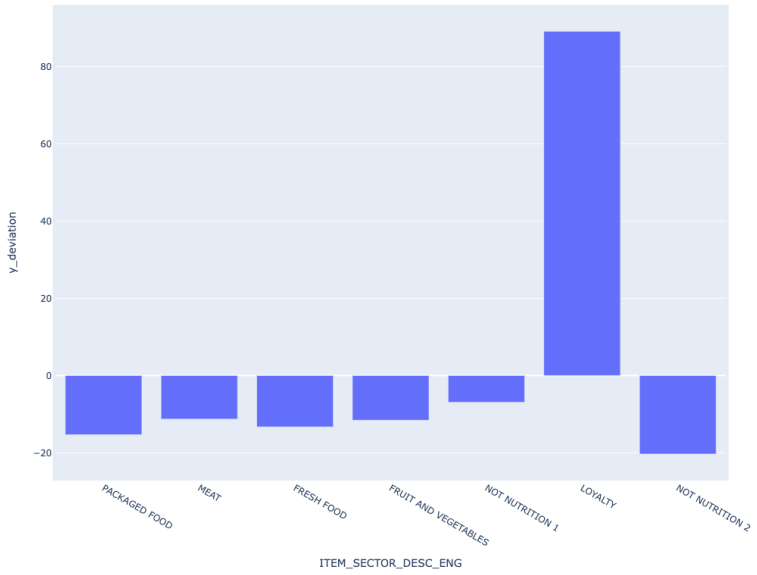

Figure 4.21: Growth rate by category of product by year (between 2020 value and baseline value) for the retail company

- Retail : most affected sectors of products are 'NOT NUTRITION 2' (i.e textiles, etc...) with a drop of 20% compared to baseline, and 'LOYALTY' with an increase of 89% compared to baseline. If we look more deeply into product categories, we can see that category 'SO: OPG' had a huge increase of sales in 2020 (need to check with business because it seems like an abnormality), category 'DF: FF TEXTILE' had a decrease of 68% in 2020 whereas 'PR: SPORTS NUTRTION' had an increase of 116% in 2020.

- Luxury : most affected category of product is 'WATCHES' with a decrease of 44% in 2020, most affected channel is 'Wholesale' with a decrease of 40% whereas E-commerce had an increase of 29%.

Store

What is the impact of COVID-19 on sales per store? Are some stores more affected than others? Now we try to identify the stores and call centers that were most affected by COVID-19. To do so, we proceed the same way as before, but this time we define a baseline per store/call center per year. We then can visualize the growth rate in 2020 compared to baseline (before 2020) for different stores/call centers.The purpose of this analysis is to understand if all stores / call centers are identically impacted by COVID-19 to identify the scope of the data processing transformation.

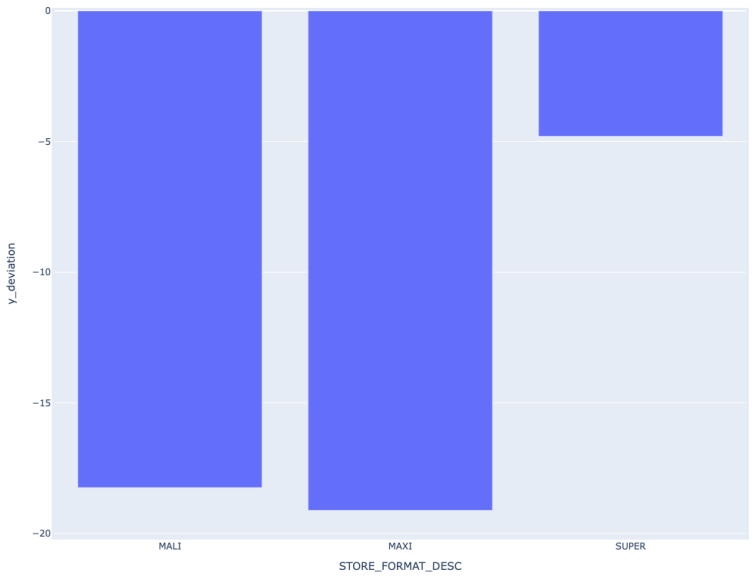

Figure 4.22: Growth rate by store by year (between 2020 value and baseline value) for the retail company

- Retail :most affected store types are 'MAXI' and 'MALI' (largest stores).
- Telecommunication : some call centers are more affected than other. Type 'D' is the category of call center with the greatest increase of calls 69%.

4.3.2 How is COVID-19 impacting sales distribution?

Seasonality

What is the impact of COVID-19 on seasonality? COVID-19 might have not only impacted volume of sales, but also its distribution. For example, some industries noticed that sales were more evenly distributed throughout the week after major lockdowns because people were working from home and therefore could more easily do shopping during the week. It is therefore necessary to check if the seasonality has changed after COVID-19 since our time series model has a seasonal component. If it has changed, we might consider changing the value of the seasonal component after COVID-19 or to change our target variable as the weekly sales volume instead of the daily sales volume.

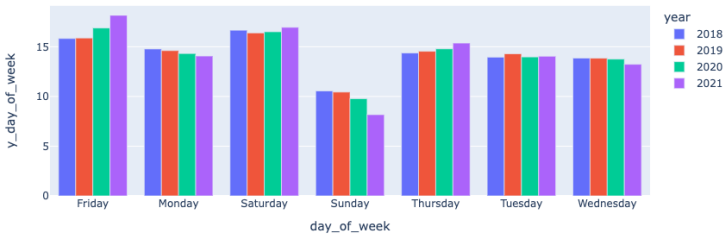

Figure 4.23: Weekly distribution of sales in quantity for the retail company

- Retail : sales seem to happen less and less on Sunday's and more and more on Friday's. A part from that there are no significant differences.
- Luxury : no clear pattern.
- Telecommunication : less and less calls on Saturday, and a bit more calls on other days of the week.

The difference of weekly distribution is negligible for all cases. We will keep our daily target variable.

What is the impact of COVID-19 on product sales distribution? (ie within a same category, are some products selling more compared to before?) Finally, we want to check if all products have been affected similarly within a same category. For example, within fresh products, are people buying more

vegan products compared to dairy products? Are people buying more online compared to wholesales ?

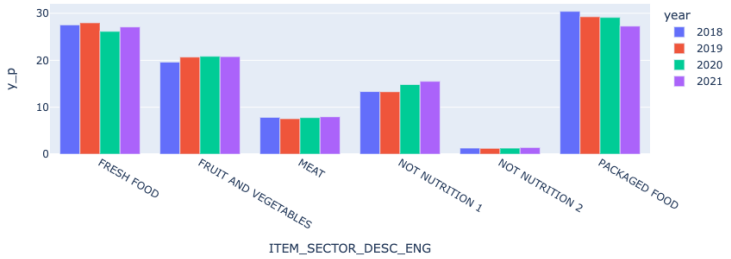

Figure 4.24: Product distribution of sales in quantity for the retail company

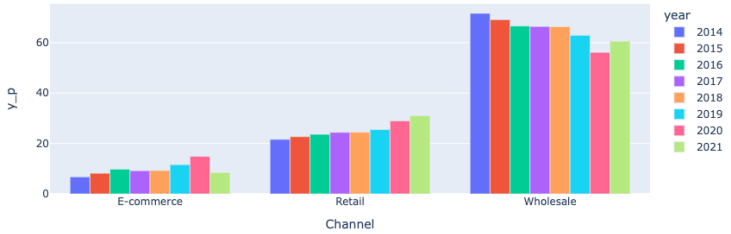

Figure 4.25: Channel distribution of sales in quantity for the luxury company

- Retail : people are putting less of their budget into PACKAGED FOOD compared to NOT NUTRITION (clothes for example).
- Luxury: less WATCHES, more LEATHER. More e-commerce and retail and less wholesale.

4.3.3 Conclusion

This EDA part's goal is to understand what is changing in 2020 due to the COVID-19 crisis and if it affects future data. We understand which scope is most affected and needs transformation. We understand if the seasonality has changed. We conclude in which scenario we are (minor impact scenario, lasting impact scenario or major disruption scenario).

For the retail company, we can consider we are in the case of a minor impact scenario : there is a limited time range where there is a significant impact of COVID-19 on sales, but things seem to get back to normal afterwards.

For the luxury company, the impact is a bit more than minor. The e-commerce channel seem to explode, whereas watches seem to be selling less and less after COVID-19. This is a lasting impact scenario.

For the telecommunication company, the time range during which there is a significant impact is not over. The future trend is difficult to estimate. It is a major disruption scenario.

Chapter 5

Increase forecast accuracy for post COVID-19 predictions

Our goal is to make post COVID-19 crisis predictions as accurate as possible. In this chapter, we will therefore explore different methods to increase forecast accuracy of post COVID-19 predictions of a time series model when there is a COVID-19 impacted set in the training data.

We switch back to our simple time series to illustrate this chapter. The univariate time series represent the sales of clothing in retail in terms of value from 1988 till 2021 in the UK. In this case, we are in a minor impact scenario and we can consider the COVID-19 period observations as anomalies.

5.1 Change the input

Modification of observations between COVID-19 start date and COVID-19 end date

5.1.1 Imputation

COVID-19 period data is replaced with previous year data The first way to correct the COVID-19 anomalies is to make the assumption that the anomaly is limited in time (March 2020-April 2021), and the new normal is actually very close to the old normal (before COVID-19). Therefore, we can try to remove the data points that are in that period and replace them with a baseline period.

- Baseline : 2018-03-01 $\rightarrow$ 2019-04-30

- Covid period : 2020-03-01 → 2021-04-30

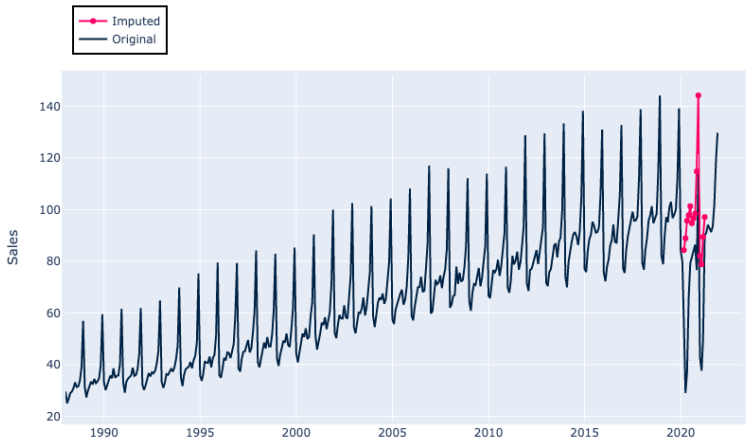

Figure 5.1: Imputed time series and original time series

5.1.2 Imputation

Covid period data is replaced with forecasted data Another imputation method could be to replace the COVID-19 period with forecasted values by a first model (for example a Prophet model) instead of replacing them with previous year data.

5.1.3 Multiplicative adjusted coefficient

Covid period data is multiplied by a coefficient Instead of replacing all data points by previous year or by forecasted values, we could multiply all data points by a multiplicative adjusted coefficient so the values look more like previous values but without replacing them.

For example, we can :

- Compute a = average yearly total sales
- Compute b = COVID-19 period yearly total sales
- Multiply all COVID-19 period data points by a/b

This way, we adjust the time series value during COVID-19 period by keeping the original shape of the time series curve.

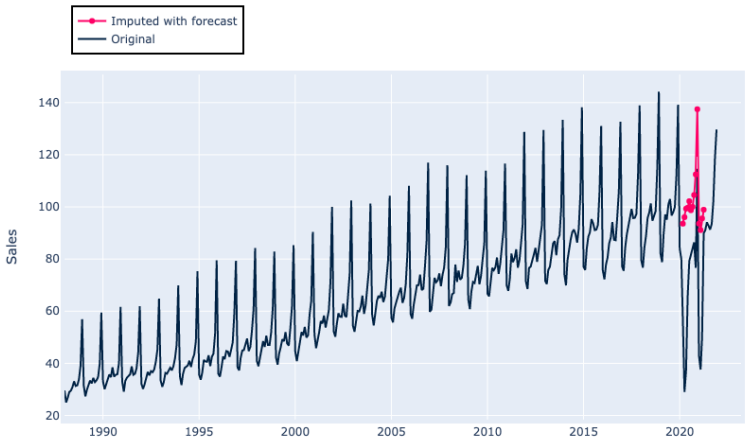

Figure 5.2: Imputed time series with forecasted values by a Prophet model and original time series

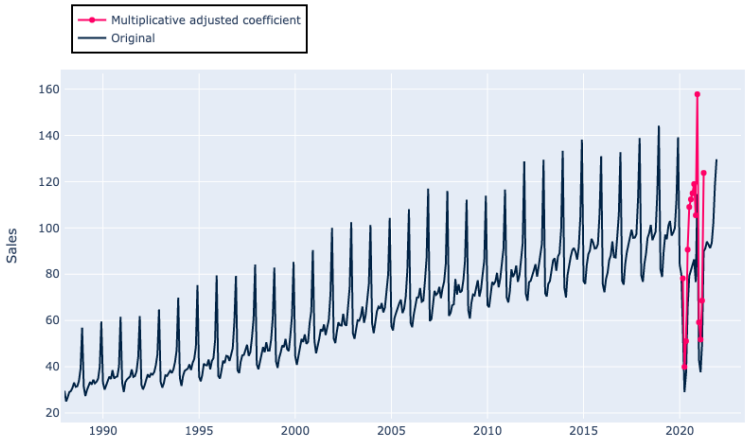

Figure 5.3: Multiplicative adjusted coefficient time series and original time series

5.1.4 Optimal transport

Covid period data is used as source distribution and transported to baseline data which is the target distribution

Why Optimal Transport?

Few have tried to use optimal transport for domain adaptation on time series. Here, we will use optimal transport to adapt sequence of observations in time considered as an anomaly to another sequence of observation in time considered as normal as it is described in this article posted by Quantmetry [21].

OT for time series context adaptation We first consider the set of abnormal observations as a $1D$ vector (the observation values) which is the source context. Then, to get rid of the seasonality, we chose as target context the normal observations from the same period of the year as the abnormal observations but a year where the observations were normal (a year before covid-19 and preferably the most representative of future observations).

So the contexts are :

- Baseline : 2018-03-01 → 2019-04-30
- Covid period : 2020-03-01 → 2021-04-30

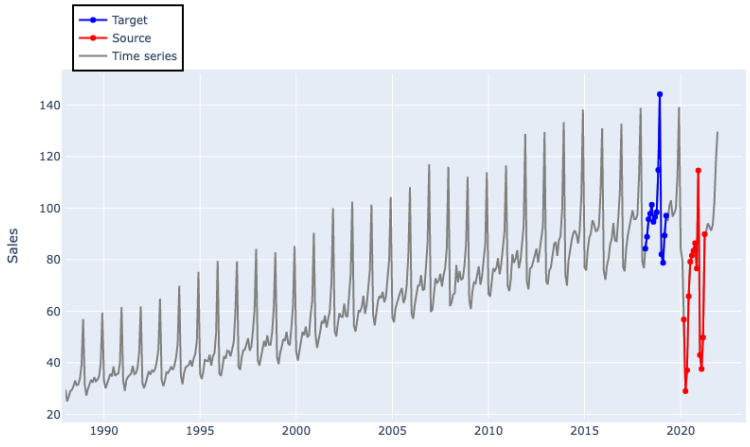

Figure 5.4: Optimal transport for simple univariate time series context adaptation (before transformation)

We then apply optimal transport using euclidean distance as cost function. Optimal transport aims then at reducing the distance between the data points from the source context (abnormal observations) and the target context (normal observations). However, for some use cases, to avoid computational limitations and assure to reach the optimal solution, we can subdivide the problem. It is the case of the luxury demand prediction problem, in which we divide the original time series into 3 categories of products x 3 channels of distribution x 2 sequence of time = 18 time series.

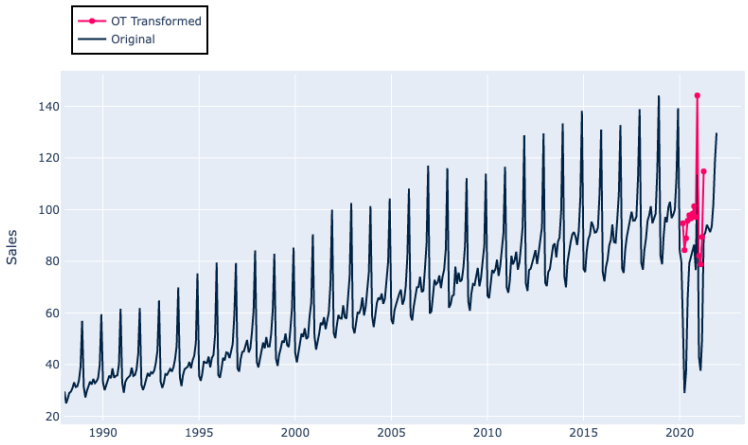

Figure 5.5: Optimal transport for simple univariate time series context adaptation (after transformation)

First time series sequence:

- Baseline : 2018-03-01 → 2018-09-30
- Covid period : 2020-03-01 → 2020-09-30

Second time series sequence:

- Baseline : 2018-10-01 → 2019-04-30
- Covid period : 2020-10-01 → 2021-04-30

5.2 Add external features to model

Instead of removing data, we introduce external regressors that can help the algorithm find correlation between the target variable and the regressor.

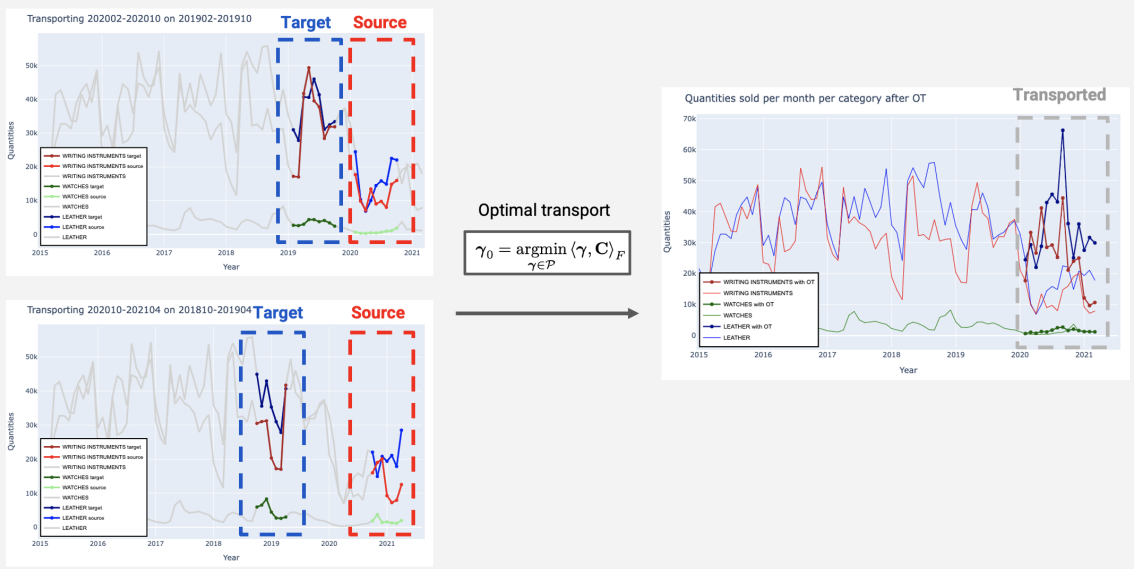

Figure 5.6: Optimal transport for complex time series context adaptation

5.2.1 Boolean feature (COVID-19 flag)

This method consists of marking the periods affected by the crisis using boolean variables, then integrating them into the model as exogenous variables. This approach allows a consideration of the disturbances without modifying the values.

We add a boolean feature that we call $covidflag$, which value is :

- 1 during COVID-19 period : 2020-03-01 → 2021-04-30
- 0 elsewhere

5.2.2 Google mobility regressor

Instead of adding a boolean feature to flag a covid impacted period, a more subtle method is to add a continuous feature that informs of the COVID-19 situation intensity. During lockdowns, people were not able to leave their homes, during moderate government regulations people were only allowed to work from home, and when there were no regulations people were able to move freely. Therefore it makes sense to use mobility as a proxy of the COVID-19 situation.We believe that mobility has been affected by the crisis and the government restrictions, and therefore mobility can be an indicator

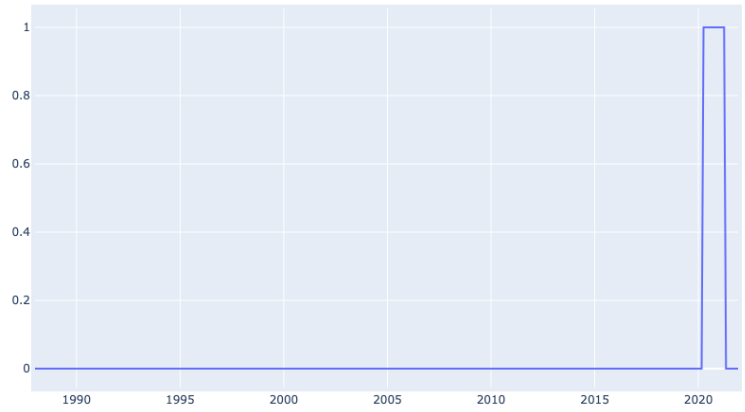

Figure 5.7: Value of boolean feature covid flag

of the COVID-19 situation. For that we use publicly available data collected by Google [25] that informs about people's mobility, available for different countries and different types of mobilities (shopping, work etc.). These datasets show how visits and different length of stay at different places change compared to a baseline. Changes for each day are compared to a baseline value for that day of the week. The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020 (just before COVID-19 global crisis). Therefore local events and seasonal changes might bias the baseline.

The data published by Google is in % change from a baseline for each region or sub-region (if present). To use Google mobility data we calculate the sum of the % reductions for all the regions of a country. We then get a percentage change from baseline for a specific country for : retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, residential. We then calculate a rolling average for every category.

To forecast future target values we need to know the future google mobility features values. However, google mobility features future values are unknown. One way to approximate google mobility features future values is to take a monthly average of that value during the 'normal' time (excluding covid period).

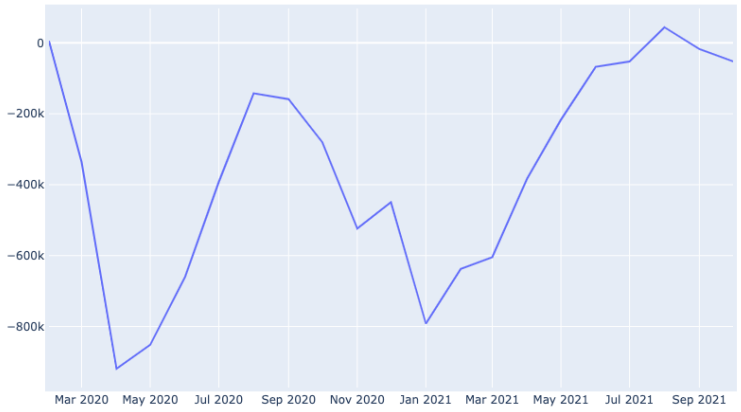

Figure 5.8: Value of Google mobility feature retail and recreation change from baseline

5.2.3 Stringency index regressor

Another continuous feature that can be added as external data is the stringency index [26]. Stringency index is a policy index that reflects government response on containment and close collected from OxCGRT Covid policy tracker. It uses 8 containment indicators plus 1 health indicator which records public information campaigns. The value of the index on any given day is the average of these 9 indicators, each taking a value between 0 and 100. The index reflects therefore several levels of government measures. The higher the index value, the stronger the government measure.

$$I = \frac{1}{9}(\sum_{i=1}^{8} C_j + H1)$$

where I is the stringency index score and I_j are the sub index score (see [26] for more information on how sub index scores are calculated).

Once again, to forecast future target values we need to know the future stringency index values. To approximate the stringency index future values is to assign an average value of the last 3 months.

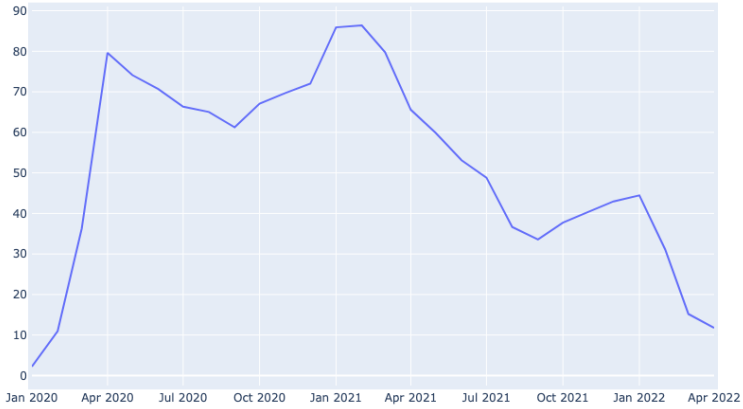

Figure 5.9: Value of Stringency index feature

Chapter 6

Results and Analysis

To compare the different methods, we calculate the forecast accuracy and the bias for each use case.

6.1 Simple univariate time series

Let's first start with the univariate time series representing the sales in clothing in retail in terms of value (amount spent) each month from January 1988 to December 2021 in the UK. We want to predict sales from July 2021 till December 2021 knowing that COVID-19 affected sales between March 2020 till April 2021. In each case, we use a Prophet model.

Table 6.1: Performance on validation set

Method	**Model**	**Forecast accuracy**	**Bias**
-	Baseline	90.3 %	−0.0887
Change the input	Imputation with past data	95.3 %	0.0138
Change the input	Imputation with predicted data	93.1 %	0.0584
Change the input	Multiplicative coefficient	94.6 %	0.0281
Change the input	Optimal transport	93.9 %	0.0399
External regressors	Boolean feature	93.3 %	0.0509
External regressors	Google mobility regressor	96.0 %	−0.0147
External regressors	Stringency index regressor	62.2 %	−0.274

In this case, what seems to work best here is adding Google mobility regressors as external data. We can see however that adding stringency index

as a feature makes the model perform badly. Most certainly because the future values of the stringency index are poorly estimated.

6.2 Complex Time series

We will now compare the same methods but a more complex time series. The time series is still sales data, but this time for a large retail company with 89 stores and 1546 different groups of products. This time we have sales data points from January 2018 till July 2021. We want to forecast May - July 2021 sales per product per store per week. For that we use a xgBoost model.

Table 6.2: Performance on validation set

Method	**Model**	**Forecast accuracy**	**Bias**
-	Baseline	64.1 %	0.0376
Change the input	Imputation with past data	71.9 %	0.112
Change the input	Imputation with predicted data	66.5 %	0.215
Change the input	Multiplicative coefficient	68.3 %	0.247
Change the input	Optimal transport	61.6 %	0.332
External regressors	Boolean feature	72.8 %	0.100
External regressors	Google mobility regressor	78.1 %	0.0171
External regressors	Stringency index regressor	73.7 %	0.0785

In this case, adding google mobility features seem to perform best both in terms of forecast accuracy and bias.

6.3 Analysis

In both previous examples, Google mobility features seem to work the best. Note that the results really depend on how well we can estimate future mobility trends.

Optimal transport seem to perform poorly on both previous examples. But it can be helpful in some situations where a transformation is needed but imputation is not the best choice. In the case of the luxury company wanting to forecast the sales of its newly released products, we applied optimal transport on the COVID-19 impacted set. The results were quite satisfying (43.6% of price weighted MAPE and -16.2 % of bias obtained with optimal transport, against a baseline of 68.6% of price weighted MAPE and -61.6 % of bias)

Chapter 7

Conclusions and future work

7.1 Conclusions

COVID-19 crisis has affected most sectors of the economy and therefore changed the patterns in the data collected by companies. In a changing world, it is even more difficult for data scientists to forecast the future when nothing as such has been observed before.

Several methods to help dealing with COVID-19 impacted datasets were described and compared. We distinguish methods that change the data (imputation, multiplicative coefficients, optimal transport) from methods that add external features (boolean feature, Google mobility features, Stringency index). We can note that all methods helped increase forecast accuracy and decrease bias almost every time.

The easier solution is to add a boolean feature to the model. It is very simple, and yet it helps the model to give better predicitons in a post COVID-19 world.

The best external feature in terms of the model's performance is Google mobility feature in our use cases because mobility is very correlated to demand. However it can be diffult to estimate future values of Google mobility features, and it is why it is not always the recommended solution.

The stringency index can help create scenarios : a worst case scenario where the stringency index is very close to COVID-19 lockdowns stringency index and a best case scenario where it is close to normal situation. This way we can estimate the impact of COVID-19 on the target variable we want to predict, and we can also help leaders analyse the scenarios and prepare in case of worst case scenarios. However, stringency index is also very hard to estimate in the future. So it is not the best method if we want to accurately

forecast future values.

Another simple solution is to perform imputation. We replace the COVID-19 impacted set by previous observed values at the same time but in a normal situation. This method can give surprisingly good results. But for some cases, like predicting newly released products, we do not have previous data. In that case this method cannot work.

Finally multiplicative coefficients or optimal transport can be used in the last case where imputation is not possible. For optimal transport we define a source domain and a target domain. This method is a bit more tricky on time series because we need to choose the good source domain and target domain, and the best method for optimization.

7.2 Limitations

Working on data COVID-19 impacted is not an easy task for a Data Scientist. The first goal of this project was to find the best method to deal with never observed events like COVID-19 that impact time series data and make the future difficult to predict, but in fact during the project I realise that such a method does not exist. Instead, it is more adapted to propose a series of different methods and chose a method for each use case. Working on three different use cases is not enough to deduce a generality.

Furthermore, not all methods have been explored in this project. Simulating number of COVID-19 cases and then exploring its relationship with the target value for predictions might be an interesting way of handling the dependencies.

Also, optimal transport was used to change the input data. However not all methods for regularization were compared. It might be interesting to include future work about this topic.

Bibliography

[1] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice (3rd ed)*. Monash University, Australia, 2021. URL: https://otexts.com/fpp3/.

[2] Chuda Dhakal. "A Naïve Approach for Comparing a Forecast Model". In: *International Journal of Thesis Projects and Dissertations (IJTPD)* (2018). URL: https://www.researchgate.net/publication/326972994_A_Naive_Approach_for_Comparing_a_Forecast_Model.

[3] William W. S. Wei. "A Naïve Approach for Comparing a Forecast Model". In: *Temple University* (2011). URL: https://www.researchgate.net/publication/266390746_Time_Series_Regression.

[4] Vibhakar Mansotra Sourabh Shastri Amardeep Sharma and Anand Sharma. "A Study on Exponential Smoothing Method for Forecasting". In: *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING* (2018). URL: https://www.researchgate.net/publication/325814694_A_Study_on_Exponential_Smoothing_Method_for_Forecasting.

[5] Vibhakar Mansotra Sourabh Shastri Amardeep Sharma and Anand Sharma. "Research on Applications of ARMA in Forecasting of Time Series". In: *Conference: 2015 International Conference on Social Science and Technology Education* (2015). URL: https://www.researchgate.net/publication/300617201_Research_on_Applications_of_ARMA_in_Forecasting_of_Time_Series.

[6] Latifa Ezzine Jamal Fattah and Zineb Aman. "Forecasting of demand using ARIMA model". In: *International Journal of Engineering Business Management* (2018). URL: https://www.researchgate.net/

publication/328633706_Forecasting_of_demand_using_ARIMA_model.

[7] Duanyang Liu and Wei Jiang. "Time Series Forecasting of temperatures using SARIMA An average Example from Nanjing". In: *IOP Conference Series Materials Science and Engineering* (2018). URL: https://www.researchgate.net/publication/326880803_Time_Series_Forecasting_of_Temperatures_using_SARIMA_An_Example_from_Nanjing.

[8] Xuan Quang Tran. "Dynamic regression models and their applications in survival and reliability analysis". In: *Université de Bordeaux* (2014). URL: https://tel.archives-ouvertes.fr/tel-01201910/document.

[9] Lorenzo Nespoli and Vasco Medici. "Multivariate Boosted Trees and Applications to Forecasting and Control". In: *CoRR* abs/2003.03835 (2020). arXiv: 2003.03835. URL: https://arxiv.org/abs/2003.03835.

[10] Slawek Smyl Grzegorz Dudek and Paweł Pełka. "Recurrent Neural Networks for Forecasting Time Series with Multiple Seasonality: A Comparative Study". In: (2022). arXiv: 2203.09170 [cs.LG]. URL: https://arxiv.org/abs/2203.09170.

[11] Everette S Gardner Jr. "Exponential smoothing: the state of the art". In: *Journal of forecasting, Navy fleet Material Support Office, Mechanicsburg, Pennsylvania U.S.A* (1985). URL: https://www.bauer.uh.edu/gardner/Exp-Sm-1985.pdf.

[12] Eddie McKenzie and Everette Gardner. "Damped trend exponential smoothing: A modelling viewpoint". In: *International Journal of Forecasting* (2010). URL: https://www.researchgate.net/publication/223667578_Damped_trend_exponential_smoothing_A_modelling_viewpoint.

[13] Peter J. Brockwell Richard A. Davis. "Introduction to Time Series and Forecasting". In: (2016). URL: https://link.springer.com/book/10.1007/978-3-319-29854-2.

[14] Atje Setiawan Abdullah. "Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah". In: *Journal of Physics Conference Series* (2021). URL: https://www.researchgate.net/publication/348359394_Extreme_gradient_boosting_XGBoost_

method_in_making_forecasting_application_and_analysis_of_USD_exchange_rates_against_rupiah.

[15] Wenyang Liu Tingyu Weng and Jun Xiao. "Supply chain sales forecasting based on lightGBM and LSTM combination model". In: *Industrial Management Data Systems* (2019). URL: https://www.researchgate.net/publication/335955033_Supply_chain_sales_forecasting_based_on_lightGBM_and_LSTM_combination_model.

[16] Xu Feng Xiaoxiao Ye Yirui Li and Chen Heng. "A Crypto Market Forecasting Method Based on Catboost Model and Bigdata". In: *IEEE* (2022). URL: https://ieeexplore.ieee.org/document/9778789.

[17] Antoine Chabert. "Forecasting Time Series in COVID-19 Days". In: *SAP analytics* (26 May 2020). URL: https://blogs.sap.com/2020/05/26/forecasting-time-series-in-covid-19-days/.

[18] Shaz Hoda et al. "Consumer Demand Modeling During COVID-19 Pandemic". In: *PwC* (3 May 2021). URL: https://arxiv.org/pdf/2105.01036.pdf.

[19] Srihari Jaganathan. "Modeling and Predicting Demand During Pandemics using Time Series Models". In: (19 january 2021). URL: https://www.linkedin.com/pulse/modeling-predicting-demand-during-pandemics-using-time-jaganathan/.

[20] Shiqi Ou et al. "Machine learning model to project the impact of COVID-19 on US motor gasoline demand". In: *nature research* (2020). URL: https://www.nature.com/articles/s41560-020-0662-1.pdf.

[21] Alexandre Willot. "Intelligence Artificielle et Data Quality : comment corriger des données historiques impactées par la Covid 19 pour améliorer la qualité des prévisions ?" In: *Quantmetry* (2021). URL: https://www.quantmetry.com/blog/intelligence-artificielle-et-data-quality-comment-corriger-des-donnees-historiques-impactees-par-la-covid-19-pour-ameliorer-la-qualite-des-previsions/.

[22] Nicolas Courty et al. “Optimal Transport for Domain Adaptation”. In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* (2016). URL: https://arxiv.org/pdf/1507.00504.pdf.

[23] Gabriel Peyré and Marco Cuturi. “Computational Optimal Transport: With Applications to Data Science”. In: *now the essence of knowledge* (2019). URL: https://www.nowpublishers.com/article/Details/MAL-073.

[24] Sean J. Taylor and Benjamin Letham. “A Crypto Market Forecasting Method Based on Catboost Model and Bigdata”. In: *PEERJ* (2017). URL: https://peerj.com/preprints/3190.pdf.

[25] “Community Mobility Reports”. In: *Google* (). URL: https://www.google.com/covid19/mobility/.

[26] Hannah Ritchie et al. “Coronavirus Pandemic (COVID-19)”. In: *Our World in Data* (2020). https://ourworldindata.org/coronavirus.

TRITA-EECS-EX-2023:864