A heuristic approach for the study of Total Demand of Commercial Vehicles

Master Ingeniería Industrial - MII

Cordinator: Alvaro Sanchez Miralles

Director: Tilak Singh

Author: Ignacio Martinez de Salinas Ureta

Anexo I

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

An Heuristic Approach for the Study of the Total

Demand of Commercial Vehicles

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2019 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Ignacio MtzUreta Fecha: …28…/ …8…/ …2019…

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Tilak Singh Fecha: …28…/ …8…/ …2019…

Anexo H

AUTHORIZATION FOR DIGITALIZATION, STORAGE AND DISSEMINATION IN THE NETWORK OF END-OF-DEGREE PROJECTS, MASTER PROJECTS, DISSERTATIONS OR BACHILLERATO REPORTS

1. Declaration of authorship and accreditation thereof.

The author Mr. /Ms. Ignacio Martínez de Salinas Ureta **HEREBY DECLARES** that he/she owns the intellectual property rights regarding the piece of work: An Heuristic Approach for the Study of the Total Demand of Commercial Vehicles that this is an original piece of work, and that he/she holds the status of author, in the sense granted by the Intellectual Property Law.

2. Subject matter and purpose of this assignment.

With the aim of disseminating the aforementioned piece of work as widely as possible using the University's Institutional Repository the author hereby **GRANTS** Comillas Pontifical University, on a royalty-free and non-exclusive basis, for the maximum legal term and with universal scope, the digitization, archiving, reproduction, distribution and public communication rights, including the right to make it electronically available, as described in the Intellectual Property Law. Transformation rights are assigned solely for the purposes described in a) of the following section.

3. Transfer and access terms

Without prejudice to the ownership of the work, which remains with its author, the transfer of rights covered by this license enables:

a) Transform it in order to adapt it to any technology suitable for sharing it online, as well as including metadata to register the piece of work and include "watermarks" or any other security or protection system.

b) Reproduce it in any digital medium in order to be included on an electronic database, including the right to reproduce and store the work on servers for the purposes of guaranteeing its security, maintaining it and preserving its format.

c) Communicate it, by default, by means of an institutional open archive, which has open and cost free online access.

d) Any other way of access (restricted, embargoed, closed) shall be explicitly requested and requires that good cause be demonstrated.

e) Assign these pieces of work a Creative Commons license by default.

f) Assign these pieces of work a HANDLE (persistent URL). by default.

4. Copyright.

The author, as the owner of a piece of work, has the right to:

a) Have his/her name clearly identified by the University as the author

b) Communicate and publish the work in the version assigned and in other subsequent versions using any medium.

c) Request that the work be withdrawn from the repository for just cause.

d) Receive reliable communication of any claims third parties may make in relation to the work and, in particular, any claims relating to its intellectual property rights.

5. Duties of the author.

The author agrees to:

a) Guarantee that the commitment undertaken by means of this official document does not infring any third party rights, regardless of whether they relate to industrial or intellectual property or any other type.

b) Guarantee that the content of the work does not infringe any third party honor, privacy or image rights.
c) Take responsibility for all claims and liability, including compensation for any damages, which may be brought against the University by third parties who believe that their rights and interests have been infringed by the assignment.
d) Take responsibility in the event that the institutions are found guilty of a rights infringement regarding the work subject to assignment.

6. Institutional Repository purposes and functioning.

The work shall be made available to the users so that they may use it in a fair and respectful way with regards to the copyright, according to the allowances given in the relevant legislation, and for study or research purposes, or any other legal use. With this aim in mind, the University undertakes the following duties and reserves the following powers:

a) The University shall inform the archive users of the permitted uses; however, it shall not guarantee or take any responsibility for any other subsequent ways the work may be used by users, which are non-compliant with the legislation in force. Any subsequent use, beyond private copying, shall require the source to be cited and authorship to be recognized, as well as the guarantee not to use it to gain commercial profit or carry out any derivative works.
b) The University shall not review the content of the works, which shall at all times fall under the exclusive responsibility of the author and it shall not be obligated to take part in lawsuits on behalf of the author in the event of any infringement of intellectual property rights deriving from storing and archiving the works. The author hereby waives any claim against the University due to any way the users may use the works that is not in keeping with the legislation in force.
c) The University shall adopt the necessary measures to safeguard the work in the future.
d) The University reserves the right to withdraw the work, after notifying the author, in sufficiently justified cases, or in the event of third party claims.

Madrid, on .....28..... of .....August....., ...

HEREBY ACCEPTS

Signed.....Ignacio MtzUreta.....

Reasons for requesting the restricted, closed or embargoed access to the work in the Institution's Repository

Resumen

1. Introduccion

En este documento, definiremos los pasos a seguir para implementar una herramienta de estudio para el departamento de ventas de Fuso. La idea es encontrar las variables que están detrás de la evolución de la demanda total de vehículos comerciales. Para este propósito, se pueden realizar diferentes estudios para encontrar un algoritmo más óptimo para llevar a cabo este estudio. Nos basaremos en las herramientas del concepto de aprendizaje automático. En ese sentido, la herramienta se basará en el uso de Python y sus bibliotecas. Desde el proceso de filtrado de documentos hasta la obtención de resultados, desarrolle las soluciones primero en Jupyter y luego cree diferentes archivos .py que se puedan compilar cuando sea necesario. Sin embargo, una parte muy importante del proyecto será obtener una solución que sea útil y necesaria para el departamento. En este sentido, veremos cómo las primeras ideas del departamento de Big Data Analytics no fueron óptimas porque la solución no era lo que ya estábamos buscando o teníamos un lugar de estudio y una aplicación que el departamento de ventas no quería llevar terminado. Esto se estudia tanto en la sección de estado del arte como en la sección de metodología. Expondremos los diferentes algoritmos que hemos estudiado como parte de la preparación de los números de datos por parte del departamento de ventas y su posterior procesamiento para que pueda ser procesado por los diferentes algoritmos.

Para estudiar los valores de esta demanda de vehículos eléctricos, el departamento de ventas tenía diferentes variables que creían que podrían ser interesantes al estudiar su relación con los datos mostrados anteriormente. Algunas de las variables que se muestran a continuación nos acompañarán durante todo el estudio, mientras que otras se perderán en el camino. Debido a la falta de fiabilidad en la fuente de la misma o total ausencia de correlación con las variables estudiadas. También por falta de datos suficientes para llevar a cabo algún algoritmo necesario para encontrar la relación mencionada anteriormente que buscaremos durante todo el proyecto. La lista de variables que se estudiaron en el primer paso del proyecto se muestra a continuación:

yo. Tierra agricola
ii) Transporte aéreo, flete (millones de toneladas-km)
iii) Batalla relacionada con muertes

iv. Costo de exportación
v. Costo de importación
vi. IPC Índice de precios al consumidor
vii. Población ocupada por situación en el empleo
viii. PIB por tipo de gasto
ix. Valor Bruto Agregado
X. Homicidio intencional
xi. Tasas de interés
xii. Población por actividad, estado, edad, sexo y residencia.

La idea, por lo tanto, será crear un documento que muestre este análisis de manera resumida y visual. Esta solución estará influenciada por el hecho de que los miembros del departamento de ventas no tienen un amplio conocimiento en programación. Por lo tanto, intentaremos encontrar una solución que tenga una implementación fácil y simple y que no incluya la modificación de ningún código interno. Es por eso que intentaremos crear una herramienta que sea lo más robusta posible para obtener el mejor tipo de documentos. Ser capaz de tratar a cada uno de una manera particular y diseñar los diferentes formatos a procesar. En este sentido, también habrá una conversación fluida con el departamento de ventas para establecer un formato óptimo para los documentos a tener en cuenta. En este sentido, nos guiaremos por los códigos ISO3 para clasificar los países y con diferentes códigos para cada variable que veremos más adelante en la metodología cuando estudiemos el proceso de limpieza y procesamiento que han sufrido los documentos para ser procesados por el programa.

También describiremos más adelante todas las herramientas utilizadas al producir esta herramienta. Como se mencionó anteriormente, la mayor parte del desarrollo de la aplicación se realizará en Python con la ayuda de los Cuadernos Jupyter. También te apoyaremos en el editor de texto Sublime Text. Eso tiene diferentes complementos para poder lidiar con diferentes tipos de código. Entre ellos Python.

Por lo tanto, el propósito de este documento se resume como la creación de una herramienta útil y visual que servirá con todos los requisitos del departamento en cuestión.

2. Objetivos

El objetivo de este proyecto es lograr un modelo justo para predecir la evolución de la curva de la demanda total de los vehículos comerciales. El modelo debe ser fácil de usar y actualizar, ya que los usuarios no serán personas técnicas y deberán poder usarlo una vez finalizado el proyecto. El modelo debe dar una idea basada en el país

seleccionado, como: las variables que explican la evolución de la demanda para ese país (o grupo que pertenece a ese país) y mostrar la evolución predicha basada en estas variables. Uno de los objetivos también sería una interfaz de usuario, pero el tiempo para completar este proyecto es de 5 meses. La interfaz de usuario descansará como un segundo objetivo.

Además, hay otro tipo de objetivos, centrados en el crecimiento de los abajo firmantes:

- Ser capaz de manejar la presión de un proyecto de este tamaño.
- Aprenda cómo funcionan los enfoques tradicionales de un pronóstico de tiempo
- Desarrollar un nuevo modelo.
- Aprenda y seleccione las tecnologías utilizadas en este nuevo modelo

3. Solucion

La solución se basaría en el estudio de correlación. Creamos algunos algoritmos para obtener la cantidad de similitud existente entre la Demanda total de vehículos comerciales y las variables.

Para esto, utilizaremos el coeficiente de Spearman para medir la cantidad de similitud que existe entre el comportamiento o la evolución de dos series de tiempo: la demanda total de vehículos comerciales para un país y las diferentes variables que estudiamos en el capítulo Clasificación. Una vez que alcanzamos algunos resultados interesantes, automatizamos y optimizamos el proceso para poder obtener los mismos resultados para diferentes grupos de países. Para este estudio preliminar utilizamos un grupo de 9 países:

- Indonesia
- Hong Kong
- China
- Vietnam
- Tailandia
- Malasia

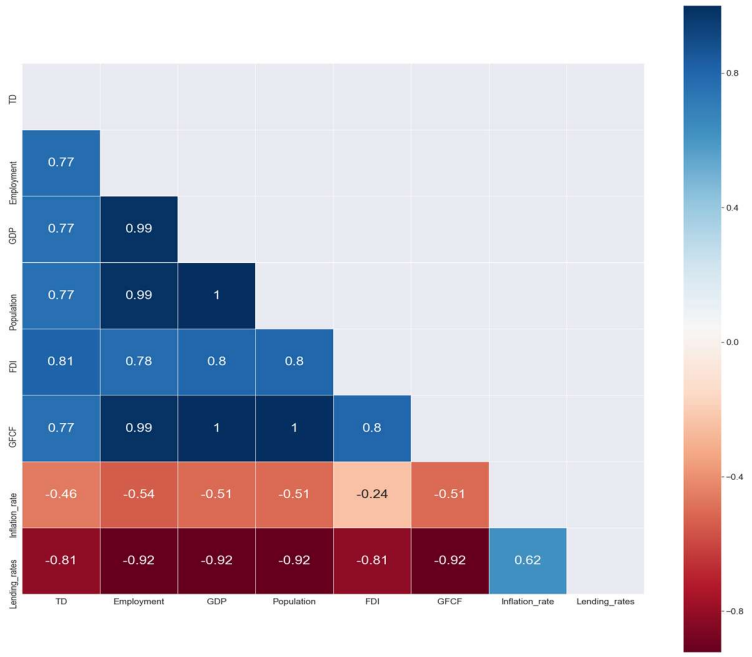
- Filipinas
- Singapur
- Taiwán

Estos países fueron seleccionados por el equipo financiero en función de la importancia de su proyecto. Desde el proyecto se centró en la evolución de los países asiáticos. Este grupo estaba resumiendo el comportamiento de los países dentro de esa región. La idea era estudiar todos los diferentes tipos de países. Desde los muy bien desarrollados como China o Hong Kong, hasta los crecientes como Indonesia hasta los que no son muy interesantes en este momento pero podrían ser interesantes en el futuro cercano como Vietnam. Comenzamos a filtrar y preparar los datos disponibles para estos países con respecto a la demanda total de vehículos comerciales. Hicimos algunos experimentos para verificar si los datos disponibles eran confiables o no.

Para simplificar esto, creamos una matriz usando todas las variables disponibles para cada país. Las variables que finalmente utilizamos para este estudio son:

- Producto Interno Bruto
- Valor Bruto Agregado
- Inversión extranjera directa
- Tasas de interés
- Empleo
- Población
- GFCF

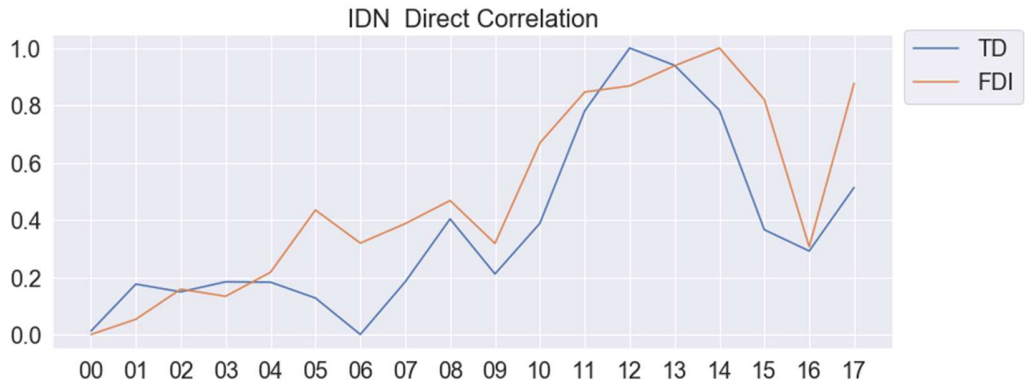
Por lo tanto, creamos una matriz con todos los coeficientes de Spearman para todas las variables y también el coeficiente para las variables entre sí. De esta manera, también podríamos obtener algunas ideas sobre cómo se comporta la estructura interna del país. Podríamos ver, por ejemplo, cómo se correlaciona el PIB con las tasas de interés. Un ejemplo de esto se puede encontrar en la siguiente figura. Si queremos centrarnos en la demanda total, deberíamos leer solo la primera columna de la matriz. Allí es donde podemos encontrar el coeficiente del lancero para todas las variables incluidas en este estudio.



La segunda parte de este enfoque fue crear una tabla que resumiera estos resultados. Esta fue una idea del equipo financiero. De esta forma, podríamos englobar todos los resultados en una tabla y cada vez que alguien necesite mirar los resultados de un país, sería fácil para ellos encontrar las variables más correlacionadas. Dentro de esta tabla también incluimos algunas variables técnicas como el valor de P. Como el equipo financiero no estaba familiarizado con esta terminología, agregamos más columnas para explicar este tipo de variables. Como podemos observar en la siguiente figura, estas dos columnas están a la derecha y se llaman: Confianza y Correlación fuerte. Agregamos algunos colores para que sea más fácil identificar las variables que tenían una fuerte correlación. En la siguiente figura podemos observar la primera versión de esta tabla.

	Total Demand	Variable_2	Spearmans Coeffcient	P_Value
1	TD	Employment	-0.091895	0.716862
2	TD	GDP	-0.057821	0.819729
3	TD	Population	-0.091895	0.716862
4	TD	FDI	-0.015488	0.951363
5	TD	GFCF	-0.065049	0.797613
6	TD	Inflation_rate	-0.524300	0.025508
7	TD	Lending_rates	0.136364	0.589515

La última parte de este enfoque fue crear dos gráficos con los resultados de la matriz de correlación. Para ver y validar que los resultados fueron correctos. Creamos dos gráficos, uno para la correlación directa y otro para la correlación inversa. De esta forma podemos ver claramente cómo la evolución en el tiempo de la demanda total de vehículos comerciales se correlaciona con las diferentes variables. Un ejemplo de la correlación directa se puede encontrar en la siguiente figura.



En este capítulo describiremos la entrega final que le dimos al equipo financiero. Como vemos todos los diferentes enfoques que intentamos

estudiar, resumiremos cuál fue el resultado final. El resultado final se basó en el análisis de correlación que explicamos en capítulos anteriores. Esto se debió a las necesidades del departamento. No avanzamos en otros enfoques, como DTW y DBA, ya que la complejidad era lo suficientemente grande y no pudimos finalizar su implementación con el tiempo que teníamos. Por lo tanto, para satisfacer las necesidades del departamento, todos estuvimos de acuerdo en que la mejor solución era dar un producto fácil y autoexplicativo. Por lo tanto, los resultados finales son un resumen del análisis de correlación, tomando los tres resultados principales de su estudio: la matriz de correlación, la tabla de correlación y las gráficas de correlación. Creamos resultados para todos los países enumerados en ese capítulo para todas las variables seleccionadas para el estudio.

Países:

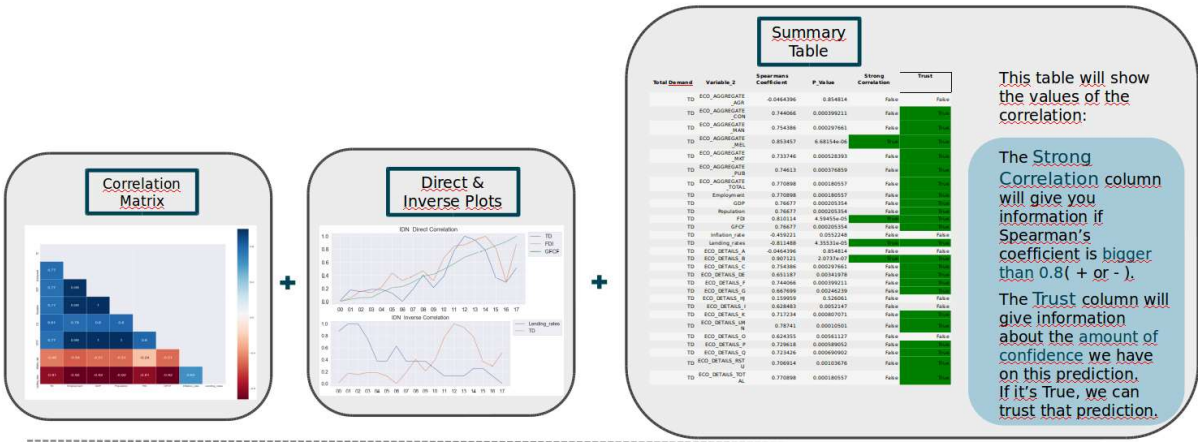
- Indonesia
- Hong Kong
- China
- Vietnam
- Tailandia
- Malasia
- Filipinas
- Singapur
- Taiwán

Variables:

- Producto Interno Bruto
- Valor Bruto Agregado
- Inversión extranjera directa
- Tasas de interés

- Empleo
- Población
- GFCF

Por lo tanto, la entrega era un documento con todas las ideas para todos estos países. Aumentamos el alcance del proyecto utilizando todos los sectores dentro de la variable Empleo. El siguiente paso fue hacer el estudio para los países con sede en África donde la confianza para los datos de la demanda total fue menor. En la siguiente figura podemos ver cómo se vería el resultado final.



Por lo tanto, el resultado final le daría al equipo financiero información rápida, información confiable y conocimiento. Esto, dentro del gran alcance del equipo financiero, debería ser realmente útil para localizar las variables más importantes con respecto a la demanda total de vehículos comerciales. Dado que el objetivo principal de este proyecto era clasificar a los países en mercados en alza o en mercados en declive, este resultado ofrece un buen punto de partida para continuar clasificando a los países con un punto de vista basado en datos.

Podemos resumir los aspectos principales de este proyecto, o conclusiones, en tres puntos principales:

- El conocimiento sobre la categorización fue adquirido por el autor.
- Se desarrolló una base sólida para futuros desarrollos.
- Finalmente se entregó un producto a los compañeros de Daimler.

Para desarrollos futuros, debemos tratar de enfocar nuestros esfuerzos en encontrar fuentes más confiables donde obtener información más útil. Para eso ya está establecida una base sólida donde, cuando se almacena la cantidad suficiente de datos, pueden tener lugar proyectos realmente interesantes.

Como conclusión personal, el proyecto fue interesante y desafiante. Haciendo un gran esfuerzo para poder manejar la importancia del proyecto.

5. Bibliografía

1.- https://www.bls.gov/cpi/

2.- https://en.wikipedia.org/wiki/Gross_domestic_product

3.- https://en.wikipedia.org/wiki/Gross_value_added

4.- https://www.investopedia.com/terms/i/interestrate.asp

5.- https://en.wikipedia.org/wiki/Principal_component_analysis

6.https://www.researchgate.net/publication/236015341_Interpretation_of_singular_spectrum_analysis_as_complete_eigenfilter_decomposition

7.- http://ssa.cf.ac.uk/zhigljavsky/pdfs/SSA/SSA_encyclopedia.pdf

8.- https://www.sciencedirect.com/science/article/pii/S0898122110003858

9.- http://www.mathcs.emory.edu/~lxiong/cs730_s13/share/slides/searching_sigkdd2012_DTW.pdf

10.- https://link.springer.com/content/pdf/10.1007/s10994-005-5828-3.pdf

11.-https://medium.com/@shachiakyaagba_41915/dynamic-time-warping-with-time-series-1f5c05fb8950

12.-https://s3.amazonaws.com/academia.edu.documents/45800856/Soheily-KhahPIATSA.pdf?response-content-disposition=inline%3B%20filename%3DProgressive_and_Iterative_Approaches_for.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20190702%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190702T081644Z&X-Amz-

Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=b2b4827faf2363b8d94d2516806cf8dcd70d62044b37c8d1d2d5780f9b7a94c7

13.- https://www.francois-petitjean.com/Research/Petitjean2014-ICDM-DTW.pdf

14 .- https://www.statisticshowto.datasciencecentral.com/autoregressive-model/

15.- https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q

Summary

1. Introduction

In this document, we will define the steps to follow to implement a study tool for the sales department of Fuso. The idea is to find the variables that are behind the evolution of the total demand of commercial vehicles. For this purpose, different studies can be carried out to find a more optimal algorithm to carry out this study. We will build on the tools of the machine learning concept. In that sense, the tool will be based on the use of Pyhton and its libraries. From the process of filtering the documents to obtaining the results, develop the solutions first in Jupyter and then create different .py files that can be compiled when necessary. However, a very important part of the project will be obtaining a solution that is useful and necessary for the department. In this sense, we will see how the first ideas of the Big Data Analytics department were not optimal because the solution was not what we are already looking for or we had a place of study and an application that the sales department did not want to carry finished. This is studied both in the State of the art section and in the methodology section. We will expose the different algorithms that we have studied as part of the preparation of the data numbers by the sales department and its subsequent processing so that you can be processed by the different algorithms.

To study the values of this demand for electric vehicles, the sales department had different variables that they believed could be interesting when studying their relationship with the data previously shown. Some of the variables shown below will accompany us throughout the study while others will be lost along the way. Due to lack of reliability in the source of the same or total absence of correlation with the variables studied. Also for lack of sufficient data to carry out some necessary algorithm to find the above mentioned relationship that we will look for throughout the whole project. The list of variables that were studied in the first step of the project is shown below:

i. Agricultural land
ii. Air transport, freight (million ton-km)
iii. Battle related deaths
iv. Cost to Export
v. Cost to Import
vi. CPI Consumer Price Index
vii. Employed population by status in employment
viii. GDP by Type of Expenditure
ix. Gross Value Added
x. Intentional Homicide
xi. Interest Rates
xii. Population by activity, status, age, sex and residence
xiii. Railways, goods transported(million ton-km)

The idea, therefore, will be to create a document that shows this analysis in a summarized and visual way. This solution will be influenced by the fact that the members of the sales department do not have extensive knowledge in programming. Therefore, we will try to find a solution that has an easy and simple implementation and that does not include the modification of any internal code. That is why we will try to make a tool that is as robust as possible in order to obtain the greatest type of documents. Being able to treat each one in a particular way and designing the different formats to be processed. In this sense, there will also be a fluid conversation with the sales department to establish an optimal format for the documents to be taken into account. In this sense we will be guided by the ISO3 codes to categorize the countries and with different codes for each variable that we will see later in the methodology when we study the process of cleaning and processing that the documents have suffered in order to be processed by the program.

We will also describe later all the tools used when producing this tool. As previously mentioned, most of the development of the application will be done in Python with the help of the Jupyter Notebooks. We will also support you in the text editor Sublime Text. That has different plugins to be able to deal with different types of code. Among them Python.

Therefore, the purpose of this document is summarized as the creation of a useful and visual tool that will serve with all the requirements of the department in question.

2. Objectives

The objective of this project is to achieve a fair model to predict the evolution of the curve of the total demand of the commercial vehicles. The model must be easy to use and to update, giving that the users won't be technical people and they must be able to use it after the project is finished. The model should give insight based on the country selected, such as: the variables that explain the demand evolution for that country (or cluster that belongs such country) and show the predicted evolution based in this variables. A user interface would be also one of the objectives but giving that the time for the completion of this project is 5 months. The user interface will rest as a second objective.

Also, there are another type of objectives, focused on the growth of the undersigned:

- Being able to handle the pressure of a project of this size
- Learn about how traditional approaches to a time forecast work
- Develop a new model
- Learn and select the technologies used in this new model

3. Solution

The solution would be based on the correlation study. We created some algorithms in order to obtain the amount of similarity there is between the Total Demand of commercial Vehicles and the variables.

For this we will use the Spearman coefficient to measure the amount of similarity there is between the behavior or evolution of two time series: the total demand of commercial vehicles for one country and the different variables we studied in the Classification chapter. Once we achieve some interesting results we automatized and optimized the process in order to be able to have the same results for different groups of countries. For this preliminary study we used a group of 9 countries:

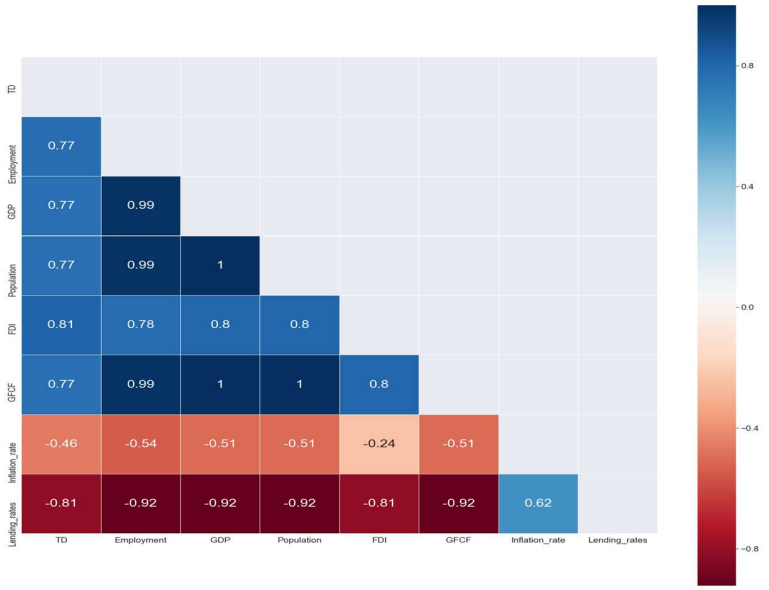
- Indonesia
- Hong Kong
- China
- Vietnam
- Thailand
- Malaysia
- Philippines
- Singapur
- Taiwan

These countries were selected by the Financial team based on the importance for their project. Since the project was focused on the evolution of the Asian countries. This group was summarizing the behavior of the countries inside that region. The idea was to study all different types of country. From the really well developed like China or Hong Kong, to the increasing ones like Indonesia to the ones that are not quite interesting right now but could be interesting in the near future like Vietnam. We started to filter and prepare the available data for these countries regarding the total demand of commercial vehicles. We did some experiments in order to check if the available data was reliable or not.

In order to simplify this we created a matrix using all the variables available for each country. The variables that we finally used for this study are:

- Gross Domestic Product
- Gross Value Added
- Direct Foreign Investment
- Interest Rates
- Employment
- Population
- GFCF

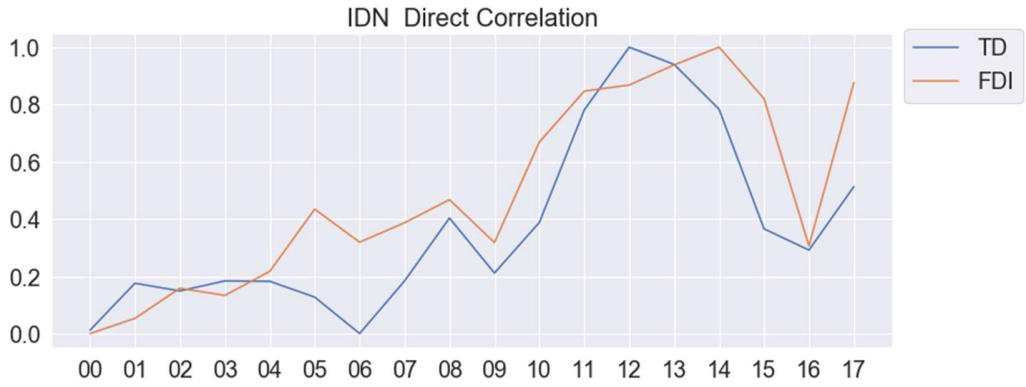
Therefore, we created a matrix with all the spearman's coefficient for all the variables and also the coefficient for the variables between each other. This way we could also get some insights about how the country internal structure is behaving. We could see for example how the GDP is correlated with the Interest Rates. An example of this can be found in the next figure. If we want to focus on the Total demand we should read only the first column of the matrix. There is where we can find the spearman's coefficient for all the variables included in this study.



The second part of this approach was to create a table that summarized these results. This was an idea of the Financial team. This way we could englobe all the results in one table and whenever someone needs to look to the results of one country it would be easy for them to find the most correlated variables. Inside this table we also included some technical variables like P value. Since the financial team wasn't familiar with this terminology we added to more columns to explain this kind of variables. As we can observe in the next figure these two columns are on the right and they are called: Trust and Strong Correlation. We added some coloring in order to be easier to identify the variables that had a strong correlation. In the next figure we can observe the first version of this table.

	Total Demand	Variable_2	Spearmans Coeffcient	P_Value
1	TD	Employment	-0.091895	0.716862
2	TD	GDP	-0.057821	0.819729
3	TD	Population	-0.091895	0.716862
4	TD	FDI	-0.015488	0.951363
5	TD	GFCF	-0.065049	0.797613
6	TD	Inflation_rate	-0.524300	0.025508
7	TD	Lending_rates	0.136364	0.589515

The last part of this approach was to create two plots with the results of the correlation matrix. In order to see and validate that the results were correct. We created two plots, one for the direct correlation and another one for the inverse correlation. This way we can clearly see how the evolution over time of the total demand of commercial vehicles is correlated with the different variables. An example of the direct correlation can be found in the next figure.



4. Results & Conclusions

In this chapter we will describe the final deliverable we gave to the Financial Team. Since we see all the different approaches we tried to study, we will summary what was the final result. The final result was based on the correlation analysis we explained in previous chapters. This was because of the needs of the department. We didn't move forward in other approaches like DTW and DBA since there complexity was large enough and we couldn't finished its implementation with the time we had. Therefore, in order to meet the department needs we all agreed that the best solution was to give an easy and self explanatory deliverable. Therefore, the final results is a summary of the correlation analysis, taking the three main results from its study: the correlation matrix, the correlation table and the correlation plots. We created results for all the countries listed in that chapter for all the variables selected for the study.

Countries:

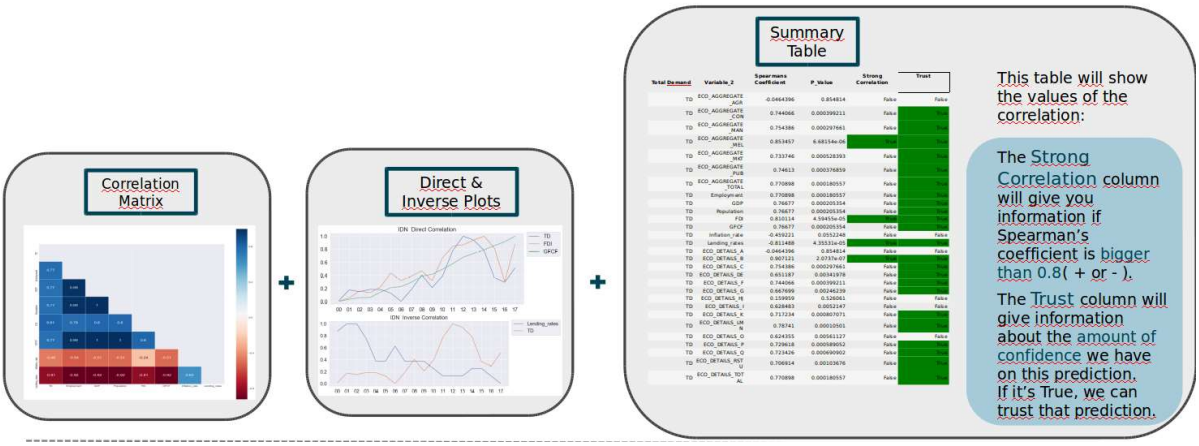
- Indonesia
- Hong Kong
- China
- Vietnam
- Thailand
- Malaysia

- Philippines
- Singapur
- Taiwan

Variables:

- Gross Domestic Product
- Gross Value Added
- Direct Foreign Investment
- Interest Rates
- Employment
- Population
- GFCF

Therefor, the deliverable was a document with all the insights for all these countries. We increased the scope of the project using all the sectors inside the Employment variable. The next step was to make the study for countries based in Africa where the confidence for the data of the total demand was smaller. In the next figure we can see how the end result would look like.



Therefor, the final result would give the Financial team fast insights, reliable information and gain of knowledge. This, inside the big scope of the financial team, should be really helpful for localizing the most important variables regarding the total demand of commercial vehicles. Given that the main goal of this project was to categorize the countries into rising markets or declining markets this results gives a

good starting point to continue to categorize the countries with a data driven point of view.

We can summarized the main aspects of this projects, or conclusions, in three main points:

- Knowledge was gained
- A solid base was developed for future endeavors
- A deliverable was finally handover to the Financial colleagues

For future developments, we should try to focus our efforts in finding more reliable sources where to get more useful information. For that is already established a solid base where, when the enough amount of data is stored, really interesting projects can take place.

As a personal conclusion, the project was interesting and challenging. Making myself to make a great effort to be able to handle the importance of the project.

5. Bibliography

1.- https://www.bls.gov/cpi/

2.- https://en.wikipedia.org/wiki/Gross_domestic_product

3.- https://en.wikipedia.org/wiki/Gross_value_added

4.- https://www.investopedia.com/terms/i/interestrate.asp

5.- https://en.wikipedia.org/wiki/Principal_component_analysis

6.https://www.researchgate.net/publication/236015341_Interpretation_of_singular_spectrum_analysis_as_complete_eigenfilter_decomposition

7.- http://ssa.cf.ac.uk/zhigljavsky/pdfs/SSA/SSA_encyclopedia.pdf

8.- https://www.sciencedirect.com/science/article/pii/S0898122110003858

9.- http://www.mathcs.emory.edu/~lxiong/cs730_s13/share/slides/searching_sigkdd2012_DTW.pdf

10.- https://link.springer.com/content/pdf/10.1007/s10994-005-5828-3.pdf

11.-https://medium.com/@shachiakyaagba_41915/dynamic-time-warping-with-time-series-1f5c05fb8950

12.-https://s3.amazonaws.com/academia.edu.documents/45800856/Soheily-KhahPIATSA.pdf?response-content-disposition=inline%3B%20filename%3DProgressive_and_Iterative_Approaches_for.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20190702%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190702T081644Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=b2b4827faf2363b8d94d2516806cf8dcd70d62044b37c8d1d2d5780f9b7a94c7

13.- https://www.francois-petitjean.com/Research/Petitjean2014-ICDM-DTW.pdf

14 .- https://www.statisticshowto.datasciencecentral.com/autoregressive-model/

15.- https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q

Masther Thesis

INDEX

1. Introduction

In this document, we will define the steps to follow to implement a study tool for the sales department of Fuso. The idea is to find the variables that are behind the evolution of the total demand of commercial vehicles. For this purpose, different studies can be carried out to find a more optimal algorithm to carry out this study. We will build on the tools of the machine learning concept. In that sense, the tool will be based on the use of Pyhton and its libraries. From the process of filtering the documents to obtaining the results, develop the solutions first in Jupyter and then create different .py files that can be compiled when necessary. However, a very important part of the project will be obtaining a solution that is useful and necessary for the department. In this sense, we will see how the first ideas of the Big Data Analytics department were not optimal because the solution was not what we are already looking for or we had a place of study and an application that the sales department did not want to carry finished. This is studied both in the State of the art section and in the methodology section. We will expose the different algorithms that we have studied as part of the preparation of the data numbers by the sales department and its subsequent processing so that you can be processed by the different algorithms.

The structure of sales is quite heterogeneous. Therefore, a large part of the project has focused on finding the optimal structure of these files. As well as the veracity of them. Arriving at a point where there was enough data, but were not entirely reliable and it was decided to restructure the data and its sources to obtain a smaller volume of data but that were more accurate. In that sense, a wide range of variables and ways of representing the data are available. Forms that will be presented below to give a clearer idea of the initial point of the project and the objective to be achieved.
The first document of total demand for commercial vehicles has a structure as shown in Table 1.

REGIONS/COUNTRIES	2005	…	2017
EUROPE	**3,156,871**	**3,178,284**	**3,388,134**
EU 28 countries + EFTA	**2,564,001**	**2,555,352**	**2,722,937**
EU 15 countries + EFTA	**2,376,384**	**2,341,547**	**2,455,125**
AUSTRIA	37,678	38,793	41,509
BELGIUM	71,413	68,313	77,570
DENMARK	62,340	69,403	63,528
…	19,960	20,973	22,234

Table 1 - Total Demand Structure

The structure is simple. Having values from 2005 to 2017. At this time of development, we also thought about the possibility of creating a Forecast using the values that were available. A possibility that was also studied later and that we will study when we see the methodology of the project.

To study the values of this demand for electric vehicles, the sales department had different variables that they believed could be interesting when studying their relationship with the data previously shown. Some of the variables shown below will accompany us throughout the study while others will be lost along the way. Due to lack of reliability in the source of the same or total absence of correlation with the variables studied. Also for lack of sufficient data to carry out some necessary algorithm to find the above mentioned relationship that we will look for throughout the whole project. The list of variables that were studied in the first step of the project is shown below:

xiv. Agricultural land
xv. Air transport, freight (million ton-km)
xvi. Battle related deaths
xvii. Cost to Export
xviii. Cost to Import
xix. CPI Consumer Price Index
xx. Employed population by status in employment
xxi. GDP by Type of Expenditure
xxii. Gross Value Added
xxiii. Intentional Homicide
xxiv. Interest Rates
xxv. Population by activity, status, age, sex and residence
xxvi. Railways, goods transported(million ton-km)

In total a list of 13 variables. Of which, some will require a brief explanation to find the existing or non-existent relationship with the demand. This explanation will be carried out in the methodology, where the filtering and cleaning process carried out for each of these documents will also be explained in order to be able to dispose of them in an easy and generic way.

As we can see in Table 1, the structure of the demand will be saved once a year. At this point, the possibility of carrying out the study with data once a month was studied. Even different variables were obtained to carry out this study. The sales department provided data with the demand broken down in months. But the lack of other variables available to study the correlation with this structure made us abandon this initiative and that we stick to one study per year. Although it is true that the quality of the study would increase with more data, being the objective of the study to find the correlation with the different variables and not to make a forecast. The amount of data available does not influence the creation of the algorithms so much since 17 data is enough to find a monotony relationship. As will be demonstrated later.

In this same line, the possibility of increasing the number of years was studied, considering that a database of demand was available from 1997 to 2027. Including 10 years of forecast. This idea was also discarded due to the lack of confidence in the forecast made and the lack of data prior to the year 2000 for most of the variables to be studied.

One key aspect we have to take in account is that we are working with the total demand of commercial vehicles. Therefore, it includes different types of vehicles:

- Cars
- Vans
- Light Duty Trucks
- Heavy Duty Trucks

We will address this problem when we see the categorization in the methodology chapter. The idea would be to divide the total demand in this clusters so we can make a good prediction based on the type of commercial vehicle that we are using. This was a big problem during the project given that there is not a globally definition to distinguish all this types of commercial vehicles. For example, on Car of one company can be classified as a van for another one. This problem also affects to the reliability of the data, since we don't always know which values the used to calculate the total demand and to categorize the different parts.

Another approach regarding this point would be calculating the evolution for the total demand and then just used the percentage of the type of car we want to study to get an approximation on how it would affect only one type of commercial vehicles. We will elaborate on this in the upcoming chapters.

The idea, therefore, will be to create a document that shows this analysis in a summarized and visual way. This solution will be influenced by the fact that the members of the sales department do not have extensive knowledge in programming. Therefore, we will try to find a solution that has an easy and simple implementation and that does not include the modification of any internal code. That is why we will try to make a tool that is as robust as possible in order to obtain the greatest type of documents. Being able to treat each one in a particular way and designing the different formats to be processed. In this sense, there will also be a fluid conversation with the sales department to establish an optimal format for the documents to be taken into account. In this sense we will be guided by the ISO3 codes to categorize the countries and with different codes for each variable that we will see later in the methodology when we study the process of cleaning and processing that the documents have suffered in order to be processed by the program.
We will also describe later all the tools used when producing this tool. As previously mentioned, most of the development of the application will be done in Python with the help of the Jupyter Notebooks. We will also support you in the text editor Sublime Text. That has different plugins to be able to deal with different types of code. Among them Python.
Therefore, the purpose of this document is summarized as the creation of a useful and visual tool that will serve with all the requirements of the department in question.

2. State of Art

We will divide the State of Art in two main parts. First we will try to assess the forecast method regarding Machine Learning and then we will study the correlation methods. Both for Time Series treatment. That way we will follow the evolution of the project at hand. First we wanted to create this forecast using the provided data but we move from that to creating a simple correlation method where the difficult part moved from the implementation of the method to the displaying of the solution.

We will first study the most common approaches for time series. This way we will give also a brief summary of all the considerations that we have to take in account when we are studying Time Series. Specific concepts will be also taken in account in this chapter. Such as I.I.D., Central Limit Theorem, etc. All of this for considering the main aspects of data treatment before starting the actual implementation of the problem. Without all this considerations, it is likely that we achieve a misleading solution or, at least, a biased solution. It is because of it that we have to take our time to study the all the main characteristics of the data that we have before going into the creation of the tool. This is not the procedure follow in the majority of the projects and in this one, this study had to be done at the same time of the implementation of the solution. This way every step in the development of the tool was a good milestone with an accuracy and confidence index to measure the performance of the project so far.

a. Forecast Algorithms
    i. Brief history

The first autoregressive-model can be brought back to the 1930s. Where an heuristic approach was taking place in order to attack the problems with the Time Series. The work of Yule and J. Walker can be the first mention of time series analysis. After this, the Moving average methodology was introduced, H.W. would introduce the ARMA model eliminating the seasonal component in mind.

The Box-Jenkins method was the next in line. This nook included the four most used tools for Time Series analysis: specification, estimation, diagnostics and forecasting. [1]

This book set the base for the future endeavors regarding the forecasting methodology. The idea always was to obtain a prediction of the future values based on the historic ones. One of the most important aspects of the forecasting method is the treatment of the noise. We have to obtain that the noise and the errors in our predictions are behaving in a seasonal way. Once we accomplish this, we can now use statistical methods that would be really useful when predicting new values like the Central Theorem Theory.

b. Correlation Algorithms

i. Brief History

The first correlation analysis was defined as a simple regression model like:

$$y = a + bx$$

In order to minimize the amount of data, since normally the data is also depending on more variables. A normal distribution or a chi squared was normally used for distributing the explanatory variable with the idea in mind of being able to explain the dependent variable. Other distributions like Bernoulli or Pascal where developed in the 19th century.

Sir Francis Galton was the responsible of developing the correlation coefficient based on human studies, being the cousin of Charles Darwin. He stated that,, the closer to 1 the coefficient is, the stronger the linear correlation is. After this, the focus move to the treatment of the errors, creating concepts like standard deviation. [2]

3. Motivation

The motivation behind this study is the possibility or understanding the behaviour of the demand so, if one of the variables that we have defined as important for that country changes we can change our position in that market. This project initially is for the Sales Department of Fuso. So the approach will be always a practical way of position ourselves in the different markets. The focus will be in the Asian countries given the origin of this project but also with the idea in mind of being able to find new countries with good opportunities. Basing this potential not in the evolution of the total demand but in the evolution of the explanatory variables. There are variables that are obviously related with the total demand, as the gas price. The motivation then lies in the interest of the department in finding the variables that are driving the total demand of commercial vehicles for each country that is in the scope of the department. The idea of this department is to select 10/20 countries where they have to make a strong position, being that having a strong position in all the markets in practical impossible.

What they have decided is to create this study to select the countries that are more interesting for them. In this part we would have to take in account the volume of the country and the variables influencing it. For example, if we have a really big market like Indonesia but the values are not reliable, making a move to have a strong position in that market can be a risk, given that maybe the market is big enough but we don't have enough data to know if the total demand is going to increase or decrease int he upcoming years. That is why we ave to make a filter and decide which countries have a good market position aswell as a reliable source of information so different types on insights can be obtained for the study of the evolution of such time series. In this study,

the current project is only a small part of this study. We are trying to complement the information they have regarding other markets and other variables.

The strong point of this project will be the accuracy and the visual strength. The accuracy because the rest of the approaches are based in the knowledge of the workers of the department. Whom have made some really interesting approaches using Excel and the Microsoft tools. Basing their solution in the same datasets that we will be working on. In this sense, the solution of this project has to be in accordance with the solutions obtained with the other tools. The idea is always to serve a complementary and reliable information. The other aspect of this tool would be the speed in showing the response or the solution. Given that once we have created the environment for the study, if we want to get some insights in a particular country we can always run the program and obtain a good solution in seconds. This could be really useful for the workers in the Sales department. They can get some insights about a particular country just before a meeting or before traveling to a country. This way, the tool would have two purposes or motivation. To be useful for the big categorization of the countries, helping the department to choose between all the countries in the Daimler Trucks Asia market the ones where they have to take a strong position but also as a useful tool in a day-to-day basis.

Being this the main motivation behind this project there is also another aspect of it. It is the involvement of the writer of this thesis in the project. Given the structure of the department, this project would be mostly under his governance. There is a personal motivation regarding making a good performance considering all the challenges this project presents. Most of the investigation part will be under his duties. Being the question of the department as a solution for a defined problem. All the aspects regarding the development are open to discussion. So it would be a big part of the project to find the actual algorithm that suits perfectly all the challenges of the project.

Therefore the motivation will be two-folded: Business and personal.

a. Business

In this case the idea in this project is to compute the real impact of that variable in the curve and, for example, the time lapse that this changes delay in taking place. This motivation is born in Daimler Trucks Asia, given the necessity of knowing, not only the future values of the demand as a traditional forecast but also the causes which provoque this changes in the evolution of the total demand of the commercial vehicles. With this information Fuso can decide which decision take in a country based on the analysis that the model we create after this project will give. The motivation then behind this project is to get insights about why the demand it's behaving the way it is behaving. So we are not likely looking for explaining or get the futures values rather than explain the historic values so we can understand the different curves in the time series. That way, the knowledge that we will give its not based in a proper forecast that can be biased but it will only give the person some information that is true because is

based of previous values and based on that, the person will make a decision regarding if the curve will be up or down. The idea is that we don't need a really precise forecast like the Supply Chain department that has to make some decisions of the volume of production based on this forecasts. Here, we want to give some quickly and fast insights so the human can make an informed decision based in actual historic value without any prediction. This is not one hundred percent true since with some approaches we would have to make some hypothesis and try to validate it. So in this type of approach we would also have some kind of prediction or bias.

b. Personal

Being a project with this important there are also personal motivations involved in the project. The undersigned would be in charge of the development of the different algorithm used to create the final solution. Is because of that, that the motivaiton to learn and experiment must be also personal. This motivation lies in the primary point that is the continuous learning curve in the proffesional life of the undersigned. With this in mind, the personal motivation to accept and work in this project lies in thefact that this project would be helpful for the development of his career. Being a technical project where an optimized solution has to be made but also a Business related project where the solution has to be presented in a concise a visual way. Also, all the work that involves talking to other departments and trying to work with them is a good learning for the immediate professional future of the undersigned. Therefore, there are some key points where this personal motivation is summarized:

- Take a better look at the general status of the commercial vehicles market.
- Learn about machine learning techniques that are required in this project.
- Decide the approach of the project.
- Work with different departments of the company, getting to know different structures and responsibilities.

This way, the undersigned will learn about technical matters while also learning about management and organization. Working closely with the Data Science and the Data Architecs of the Big Data Team and also working with the Bussiness team trying to find the best solution for both parties.

4. Objectives

The objective of this project is to achieve a fair model to predict the evolution of the curve of the total demand of the commercial vehicles. The model must be easy to use and to update, giving that the users won't be technical people and they must be able to use it after the project is finished. The model should give insight based on the country selected, such as: the variables that explain the demand evolution for that country (or cluster that belongs such country) and show the predicted evolution based in this variables. A user interface would be also one of the objectives but giving that

the time for the completion of this project is 5 months. The user interface will rest as a second objective.

Also, there are another type of objectives, focused on the growth of the undersigned:

- Being able to handle the pressure of a project of this size
- Learn about how traditional approaches to a time forecast work
- Develop a new model
- Learn and select the technologies used in this new model

5. Methodology
    a. Introduction

We start now with the methodology that has been followed during this project. We will try to organize the steps in a more understandable way. Given that the process followed by this project wasn't always the best. We were limited by the requirements of different departments and, for example, the data that we have to study has changed multiple times over the duration of the project. This is way we will try to organize this methodology to show a rational evolution in the process of development. Taking in account that if somewhere there is a chapter that is not followed by its precedent study, this would be the primary reason for that. Also, different types of algorithm had been take in account. We will explain the method and the theory behind it in each section. We will try to summarize the key aspects of each method so we can see the objective after trying to develop it and also the reasoning after the model was implemented behind the decision taken by the Business department. Either it was to continue with the project or to pause it and continue in other directions.

    b. Cleaning and Structuring

The first part of the project involve studying the data that the Business department had provided. Here, we will have a list of documents that the department thought they could be interesting to take in account in the process of studying the evolution of the total demand. A continuation we will show a little description of each document with the objective in mind of studying the different types of data that has to merge in the next steps of the project.

        i. Agricultural land

The structure of this document is going to be the most common in this Project. Also, it is important to remark the ISO3 code for the countries. Which was a serious problem with our collaboration with the Business department and finally we optimized the change from

the official country name to the ISO3 code. Saving a lot of time and also making easier to work with the different departments involved.

Country Code	Indicator Name	Indicator Code	1960	…	1965
ABW	Agricultural land (sq. km)	AG.LND.AGRI.K2	20		20
AFG	Agricultural land (sq. km)	AG.LND.AGRI.K2	377000		378750
AGO	Agricultural land (sq. km)	AG.LND.AGRI.K2	571700		572700
ALB	Agricultural land (sq. km)	AG.LND.AGRI.K2	12320		12370
AND	Agricultural land (sq. km)	AG.LND.AGRI.K2	260		260
ARB	Agricultural land (sq. km)	AG.LND.AGRI.K2	3127460		3130980
ARE	Agricultural land (sq. km)	AG.LND.AGRI.K2	2080		2090
ARG	Agricultural land (sq. km)	AG.LND.AGRI.K2	1378290		1317800

Table 2 - Agricultural Land

ii. Air transport, freight (million ton-km)

This is the second main structure in our datasets. This type of document would require more coding in order to solver the problematic structure. This process would be automatized so we can use it with all the files that have a similar structure.

Country or Area	Year	Value
Afghanistan	2017	25.14421
Afghanistan	2016	29.01088
Afghanistan	2015	33.10204
Afghanistan	2014	34.28347
Afghanistan	2013	84.62122
Afghanistan	2012	116.6607
Afghanistan	2011	109.4211
Afghanistan	2010	108.0195
Afghanistan	2000	7.813
Afghanistan	1999	7.4
Afghanistan	1998	15
Afghanistan	1997	35.4
Afghanistan	1996	13.5
Afghanistan	1995	12.9
Afghanistan	1994	12.3
Afghanistan	1993	6.7
Afghanistan	1992	8.4
Afghanistan	1991	8.4

Afghanistan	1990	9.4

Table 3 - Air transport, freight

iii. Battle related deaths

Same structure as the previous one.

Country or Area	Year	Value
Dem. Rep. Congo	2017	1933
Dem. Rep. Congo	2016	425
Dem. Rep. Congo	2015	197
Dem. Rep. Congo	2014	985
Dem. Rep. Congo	2013	1531
Dem. Rep. Congo	…	…
Dem. Rep. Congo	2009	1978
Dem. Rep. Congo	…	…
Dem. Rep. Congo	2000	1473
Dem. Rep. Congo	1999	3282
Dem. Rep. Congo	1998	3356

Table 4 - Battle related deaths

iv. Cost to Export

This document had a really good potential as an important variable for explaining the demand. We will see in the next chapter the evolution of this dataset and the next one: Cost to Import.

Country or Area	Year	Value
Afghanistan	2014	5045
Afghanistan	2013	4645
Afghanistan	2012	3545
Afghanistan	2011	3545
Afghanistan	2010	3545
Afghanistan	2009	3030
Afghanistan	2008	2680
Afghanistan	2007	2180
Afghanistan	2006	2180
Afghanistan	2005	2180

Table 5 - Cost to Export

v. Cost to Import

Same structure as the previous one.

vi. CPI Consumer Price Index

The actual definition of CPI is:

A **Consumer Price Index** measures changes in the price level of market basket of consumer goods and services purchased by households[1].

Its formula is shown in the next Figure:

$$\text{Corsumer Price Index} = \frac{\text{Market Basket of Desired Year}}{\text{Market Basket of Base Year}} \times 100$$

Figure 1 - CPI Formula

This index is mainly used for measuring the wealth of the average population. More precisely it is used for calculating the purchasing power of the average citizen. It can be used as a good indicator on how the total demand of commercial vehicle would behave. We have to consider that inside the total demand we have cars but also vans, light duty trucks and heavy duty trucks, so this variable would be a good indicator to measure the evolution of the car demand but not so much for the evolution of the rest of the variables included in the total demand of commercial vehicles. Therefore, we would have either to compute its effect in the total demand of commercial vehicles bases on the effect on the car demand or compute a different study for commercial cars using this variable. As we can see in Table 6, the structure of this document is different. Being the key difference the frequency of the values. This variable is recorded or calculated quarterly monthly.

Country	2013	2013Q1	2013M01	2013Q2	2013M04
Afghanistan	127.80	125.22	125.19	126.72	127.00
Albania	107.58	108.46	107.57	107.94	108.91
Algeria	117.52	117.77	117.69	117.35	117.63
Angola	136.13	132.31	131.30	135.11	134.05
Anguilla	106.36	106.07	...	106.63	...
Antigua &Barbuda	108.08	108.14	108.23	108.07	107.95

Armenia, Republic of	116.80	116.15	116.30	116.16	116.21

Table 6 - CPI

The importance of this variable to measure the state of a country is quite big, it can be used to calculate the impact of the inflation as well as the value of the salaries. We will try to compute this importance in our future models.

vii. Employed population by status in employment

This document has a different structure. We would study the different variables that are used as filter inside the CSV file. The actual structure of this document is shown in Table 7.

Country or Area	Year	Sex	Status in employment	Industry	Source Year	Value
Albania	2011	Both Sexes	Total	Section A)	2014	176745
Albania	2011	Both Sexes	Total	Section B)	2014	6797
Albania	2011	Both Sexes	Total	Section C)	2014	59626
Albania	2011	Both Sexes	Total	Section D)	2014	6780
Albania	2011	Both Sexes	Total	Section E)	2014	8355
Albania	2011	Both Sexes	Total	Section F)	2014	56688
Albania	2011	Both Sexes	Total	Section G)	2014	87816

Table 7 - Employed Population

With this table we can create a lot of new tables if we decided which filter we want to use. It is because of this that we would then study the variables inside each of the headers that can be used as filters for the creation of new documents:

- Sex: States the difference between male, female and both.
- Status in employment: we have different types of contract of employment. For this initial study we will use the Total. We could create new studies based on the other filters:

        - Total
        - Contributing family members
        - Employee
        - Employer
        - Own account member
    - Industry: this filter would be very important for our study, given that the type of industry is an important variable to take in account if we have to study the variability of the total demand. We would create a study later based on the number of employment by sector of the industry:
        - Wholesale and retail trade; repair of motor vehicles and motorcycles (ISIC Rev.4: Section G)
        - Water supply; sewerage, waste management and remediation activities (ISIC Rev.4: Section E)
        - Transportation and storage (ISIC Rev.4: Section H)
        - Real estate activities (ISIC Rev.4: Section L)
        - Public administration and defence; compulsory social security (ISIC Rev.4: Section O)
        - Professional, scientific and technical activities (ISIC Rev.4: Section M)
        - Information and communication (ISIC Rev.4: Section J)
        - Human health and social work activities (ISIC Rev.4: Section Q)
        - Agriculture, forestry and fishing (ISIC Rev.4: Section A)

viii. GDP by Type of Expenditure

The Gross Domestic Product is a monetary measure of the market value of all the final goods and services produced in a specific period, often annually[2]. Similar to CPI, this value would help us to see the actual situation of a country. By studying its precedent values we can also predict the future behaviour of the country. We will use this index to correlate its evolution with the evolution of the total demand of commercial vehicles. The structure of the document can be seen in Table 8.

Country or Area	Year	Item	Value
Afghanistan	2017	Final consumption expenditure	3.51E+10
Afghanistan	2017	Household consumption expenditure (including Non-profit institutions serving households)	3.16E+10
Afghanistan	2017	General government final consumption expenditure	2.85E+09
Afghanistan	2017	Gross capital formation	2.24E+09
Afghanistan	2017	Gross fixed capital formation (including Acquisitions less disposals of valuables)	2.24E+09
Afghanistan	2017	Exports of goods and services	7.6E+08
Afghanistan	2017	Imports of goods and services	8.95E+09
Afghanistan	2017	Gross Domestic Product (GDP)	2.22E+10

Table 8 - GDP by Type of Expenditure

We can also make some remarks about the Item filter. We would study the different types of GDP we can get from this document. For simplicity, in the first study of this approach, we will use the total GDP for each country each year.

- Item:
  - Final Consumption expenditure
  - Household consumption expenditure
  - General government final consumption expenditure
  - Gross fixed capital formation
  - Export of goods and services
  - Imports of goods and services
  - Gross Domestic Product

ix. Gross Value Added (GVA)

The Gross Value added is similar to the GDP but it measures the different value added in each step of the production, for each sector. It's also a measure of the actual value of a product, taking in account the price of the product and all the cost that have been taken in account for its production[3].

Figure 2 - GVA Explained

It is therefore, a good variable to take in account for the correlation. In would make sense if the total demand of commercial vehicles is correlated with the GVA of a country given that part of the numbers to create the GVA are from the production of commercial vehicles. Even if we can get some insights for this correlation given that relation, we can still get some insights if we are going in the good direction, if our hypothesis are correct or, if we are doing something wrong since there is no correlation between some variables where it should be. The structure of this variables is similar to the GDP with the same sector taken in account.

x. Intentional Homicide

This variable was decided by the business department. This is used as a part of an approach of taking in account variables that normally the wouldn't have anything in common with the studied variable but that, maybe, the model can find some kind of correlation that the human eye can't find yet.

Country Name	1995	2000	2001	2002	2003	2004	2005
Aruba			4.305798	5.263601	4.122989	2.025583	5.998141
Afghanistan							
Angola							
Albania	6.685552	4.196069	7.045844	6.925232	5.335896	4.22888	5.001333
Andorra						1.311579	
Arab World							

United Arab Emirates					1.202587	0.660481	1.222824
Argentina	7.81233						
Armenia	3.598938	2.964567	2.884626	2.274303	2.518386	2.766102	1.945487
American Samoa			1.718951	11.91875	5.074682	6.74946	10.14919
Antigua and Barbuda		5.982006	8.229775	5.796026	5.727836	4.532218	3.361232

Table 9 - Intentional Homicide

xi. Interest Rates

This variable would be interesting for our project. The interest rates are defined as the amount a lender charges for the use of assets expressed as a percentage of the principal. The interest rate is typically noted on an annual basis known as the annual percentage rate (APR)[4]. We would see there is a strong correlation between the interest rates of the banks and the sales of commercial cars, this makes a lot of sense but we would have the problem of the categorization again. Having to distinguish between cars, vans and duty trucks.

Country	2013	2013Q1	2013M01	2013M02	2013M03	2013Q2
Afghanistan, Islamic Republic of	...	...	...	...	...	...
Albania	3.00	3.75	3.75	3.75	3.75	3.75
Angola	9.75	10.00	10.00	10.00	10.00	10.00
Armenia, Republic of	7.75	8.00	8.00	8.00	8.00	8.00
Australia	2.50	3.00	3.00	3.00	3.00	2.75
Azerbaijan, Republic of	4.75	4.75	5.00	4.75	4.75	4.75

Table 10 - Interest Rates

xii. Population by activity, status, age, sex and residence

For this variable we have different types of census and different types of reliability of the values. We will use this variable also as a validation that we are going in the right direction since it would be obvious that the total demand of commercial vehicles would be correlated with the population.

Country or Area	Year	Age	Activity status	Record Type	Reliability	Source Year	Value
Albania	2011	15 - 19	Total population	Census - de jure - complete tabulation	Final figure, complete	2014	268746
Albania	2011	15 - 19	Total economically active population	Census - de jure - complete tabulation	Final figure, complete	2014	42503

Albania	2011	15 - 19	Employed	Census - de jure - complete tabulation	Final figure, complete	2014	15854
Albania	2011	15 - 19	Unemployed	Census - de jure - complete tabulation	Final figure, complete	2014	26649
Albania	2011	15 - 19	Not economically active population	Census - de jure - complete tabulation	Final figure, complete	2014	226243
Albania	2011	15 +	Total population	Census - de jure - complete tabulation	Final figure, complete	2014	2221572
Albania	2011	15 +	Total economically active population	Census - de jure - complete tabulation	Final figure, complete	2014	958255
Albania	2011	15 +	Employed	Census - de jure - complete tabulation	Final figure, complete	2014	677950
Albania	2011	15 +	Unemployed	Census - de jure - complete tabulation	Final figure, complete	2014	280305
Albania	2011	15 +	Not economically active population	Census - de jure - complete tabulation	Final figure, complete	2014	1263317
Albania	2011	20 - 24	Total population	Census - de jure - complete tabulation	Final figure, complete	2014	243645
Albania	2011	20 - 24	Total economically active population	Census - de jure - complete tabulation	Final figure, complete	2014	106752

Table 11 - Population by activity

xiii. Railways, goods transported(million ton-km)

This variable would be interesting in a way that is correlated with the import, export data. If we found a big difference between this values and the ones in the other documents we can be sure that there is something that is not going correctly in our assumptions. We would have to create this kind of relation between variables in order to validate our hyphothesis.

The structure is similar as Cost to Export, Cost to Import.

Finally, we have another document where we have summarized all the sources of all the documents, this way we know is the data is reliable or not.

Topic	Source
Population history and projection	United Nations
Total demand	OICA
Employed population by industry	United Nations
Employed population by rural/urban area	United Nations
GDP by type of expenditure	United Nations

Gross value added by kind of economic activity	United Nations
Air transport (freight)	United Nations (WB)
Battle-related deaths	United Nations (WB)
Cost to import (USD per container)	United Nations (WB)
Cost to export (USD per container)	United Nations (WB)
Railways, goods transported	United Nations (WB)
Consumer Price Index	IMF
Monetary policy related interest rates	IMF
Agricultural land	WB
Intentional homicides	WB
UIO (CV)	OICA

Table 12 - Data Sources

c. First models – Classification

In this part we would study the different approaches that were taken in account at the start of the project. It is a good start to show the different ideas we had when the project started in order to take care of all the needs of the business department. We will see that we try different approaches, that maybe they don't have anything in common. Apart from being all from this initial part of the study.

- Location Matrix:
  One of the first ideas Tilak had was to study the countries based on the relation between its volume ad its variation. We created a couple of plot to show this difference between all the countries but it was impossible to see anything valuable:

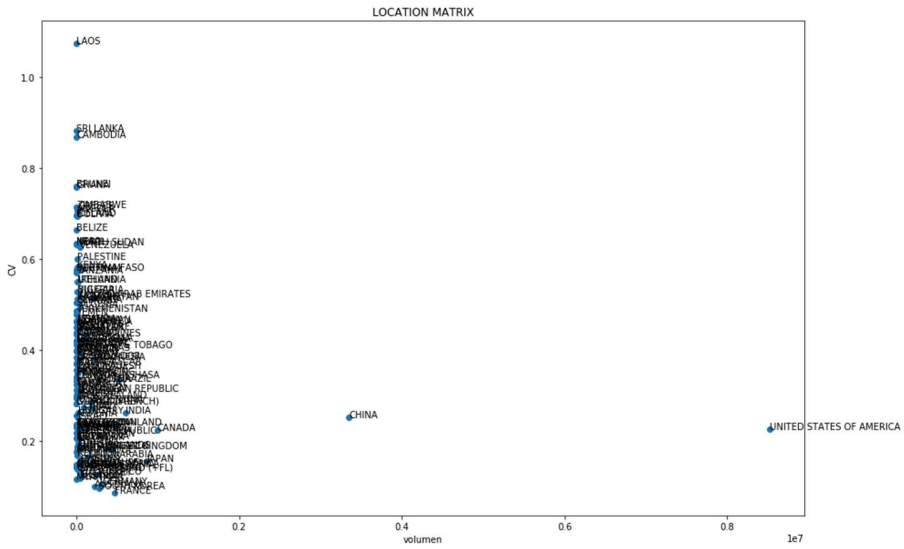


Figure 3 - Location Matrix

Even if we delete China and the United States of America from this plot in order to see something we won't get a good idea of the situation of most of the countries:

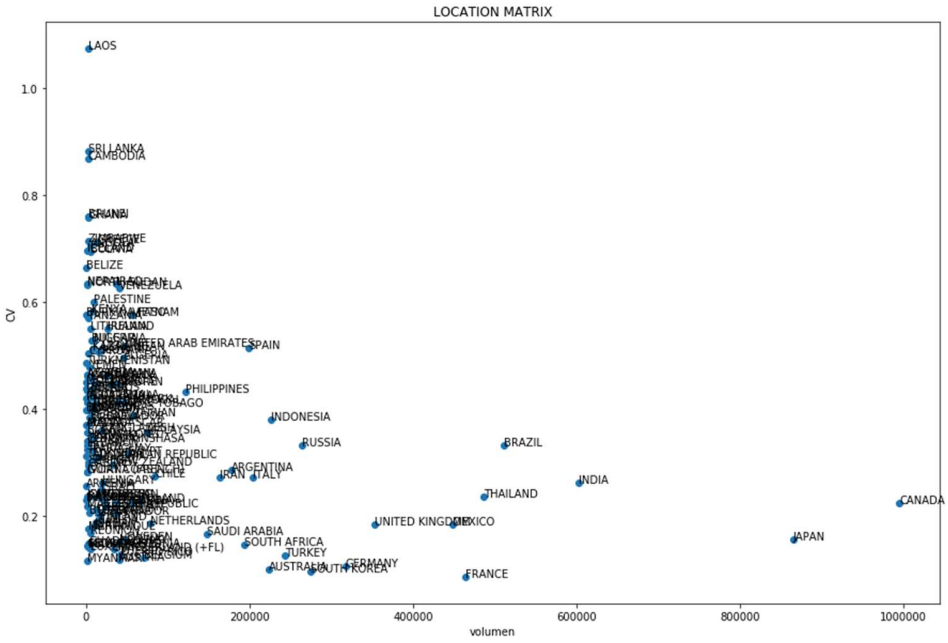


Figure 4 - Location Matrix without CHINA & US

Therefore we created the next figure in order to locate only the big countries that were defined inside our document with the Total Demand of Commercial Vehicles, this are:

- Europe

- Russia, Turkey & Other Europe
- America
- Nafta
- Central & South America
- Asia, Oceania & Middle East
- Africa

Plotting only this groups would allow us to see which group can be the most interesting to study. This would be a decision regarding Volume and the Variation considered. For this last plot we have considered the Standard Deviation as a way of calculating the variation of the total demand for each group. This was calculated using the method "std" that the pandas library has. We will see all this method in the resources chapter. The final plot can be seen in the Figure 5.

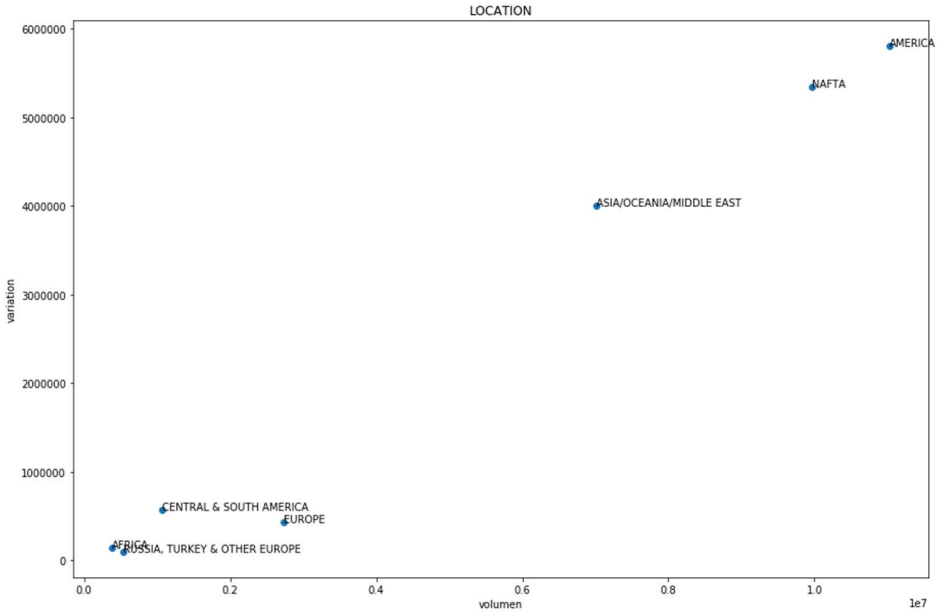


Figure 5 - Location Matrix Big Groups

We can observe here how the most interesting group to study should be America or Nafta. But considering the reduced amount of countries that we have inside of that groups we are more inclined to decide as our group of study ASIA/OCEANIA/MIDDLE EAST. This results is quite robust given that we are making this study for Daimler Trucks Asia, therefore it is a good solution that we have found that the 3rd group that has a more interesting future is the one that we need to study thoroughly.

This plot could be also shown in a different way, less intuitive but also correct. This approach can be found in Figure 6.

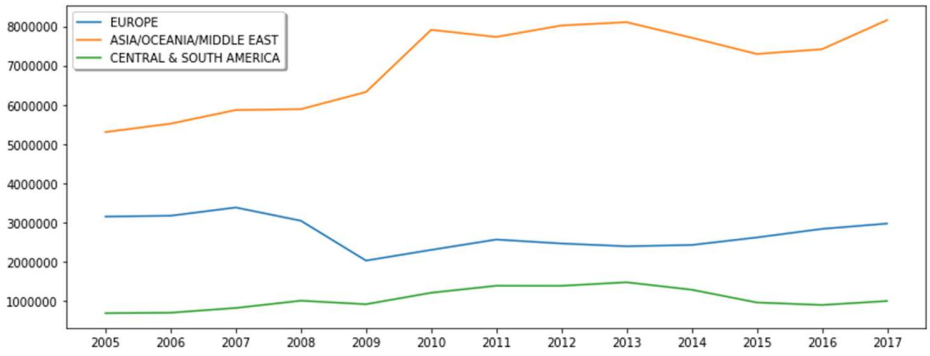


Figure 6 - Comparison Big Groups

And more precisely:

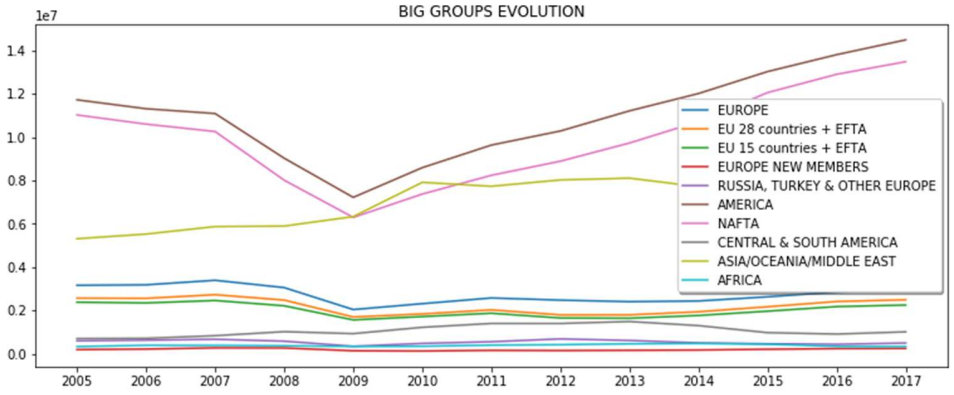


Figure 7 - Big Groups Evolution

We will rest this chose for now. Keeping in mind that that will be our last choice once we have decided the approach to attack the categorization problem. Since we don't know yet which would be the final categorization for our data we will continue with some methods considering all countries. In order to study other possibilities once the Asia group is finally assessed and we can move its methodology to other groups.

d. Data Merging, PCA and Kmeans

In this chapter we will study different possibilities for all the data that we are using. For that matter, we will first merge all the variables were we have enough data to create two machine learning methods: PCA and Kmeans.

- PCA: Principal Components Analysis

  The PCA method consists in a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly

correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components[5].
The idea of this method is to reduce the dimensions that are present in our datasets. For example, the data frame that we have after merging all the variables would have 8 dimensions:

- Total Demand
- PCI
- Interest rates
- Import Cost
- Export Cost
- GVA
- GDP

Since the volume of each variable is different we would have to normalize all the values inside our merge dataset in order to be able to use the PCA method. For this we would use PCA method from Sklearn library and Numpy. Since this library would do all the hard work, we would try to explain the theory behind this method before showing the results for our dataset.

We have already explained that we have to standardize the data set in order to have all the values between the same range of values, in this case 0 & 1. That way we can start with the classification of each value in of Principal Component. For the standardization we would use the method Standard Scaler of the library Sklearn. The math between this methods is the usual after standardization. It can be find in Figure 8.

$$z = \frac{value - mean}{standard\ deviation}$$

Figure 8 - Standardize value

Once we can compare the data, we should create a covariance matrix. This matrix is used to see the correlation between the different inputs. We have the normal considerations we have to take in account with a covariance matrix:

$$(\mathrm{Cov}(a,a) = \mathrm{Var}(a))$$

Figure 9 - Covariance Matrix Characteristic (1)

Since the covariance of the value in the position ii would be de correlation between a variable with itself we can compute the variance at that position.

$$(\mathrm{Cov}(a,b)=\mathrm{Cov}(b,a))$$

Figure 10 – Covariance Matrix Characteristics (2)

Since we are comparing a variable with another, the order doesn't matter. This is the reason why we get a symmetric matrix as a covariance matrix. Next step would be to calculate the eigenvector of this matrix in order to minimize the dimension used to explain the same data. The eigenvectors would be a linear representation of the majority of variables possible. This way, all the data would be represented by our Principal Components but in a biased or inaccurate way. With PCA method we gain simplicity and speed of calculation but we lose a lot of accuracy since what we are doing is computing a couple of variables into one. Every time we do this computation we will be losing some information.

Since the PCA method of the Sklearn library lets us decide the number of Principal Components to be computed we would decide to create 5 Principal Components. This way we won't lose a lot of information in the formation of the eigenvectors but we would have had decrease the complexity of our values by 3 degrees.

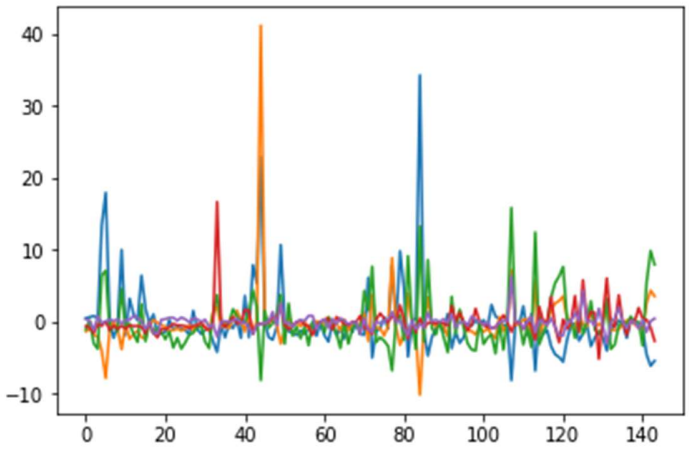


Figure 11 - PCA Analysis

This is the final result of our PCA study. We will try to continue with this method in the next iterations of our model.

- Kmeans:

  The first idea after this method was to be able to visualize data with needing to filter the country that we were looking for. Since the Kmeans doesn't work perfectly with time series, we only used this method to display the evolution of different countries. This way we could get some insights about some particular countries in a very fast way. We only had to cluster the data and then search fo the country in all the cluster, once the country was found we would plot that cluster with the country in question compared with other countries that had a similar volume. That way we could get a fast idea of how the country we were studying was doing. We can see in Figure 12 an example of this clustering.

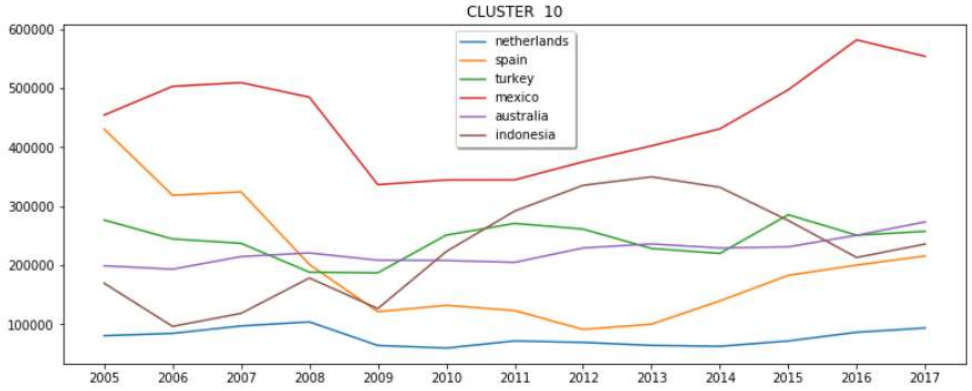


Figure 12 - Kmeans clustering

e. Correlation Study

Once we finished with the previous method. We moved forward to one of the most interesting approaches for the Financial team. As a result, this approach is the one that will be explained further in the Results and Conclusion chapter since is the one we finally delivered to the Financial team. The idea is to find the variables that are more correlated to the total demand. For this we will use the Spearman coefficient to measure the amount of similarity there is between the behavior or evolution of two time series: the total demand of commercial vehicles for one country and the different variables we studied in the Classification chapter. Once we achieve some interesting results we automatized and optimized the process in order to be able to have the same results for different groups of countries. For this preliminary study we used a group of 9 countries:

  - Indonesia

- Hong Kong
- China
- Vietnam
- Thailand
- Malaysia
- Philippines
- Singapur
- Taiwan

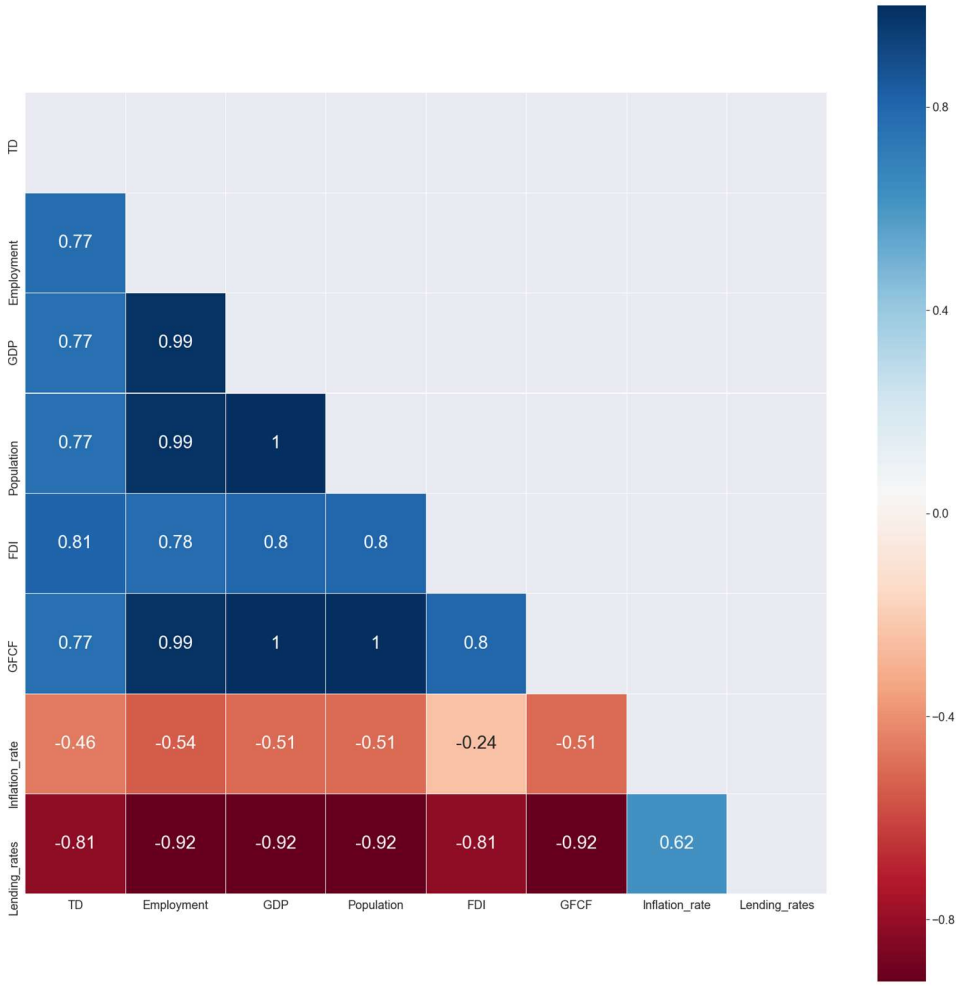
These countries were selected by the Financial team based on the importance for their project. Since the project was focused on the evolution of the Asian countries. This group was summarizing the behavior of the countries inside that region. The idea was to study all different types of country. From the really well developed like China or Hong Kong, to the increasing ones like Indonesia to the ones that are not quite interesting right now but could be interesting in the near future like Vietnam. We started to filter and prepare the available data for these countries regarding the total demand of commercial vehicles. We did some experiments in order to check if the available data was reliable or not. We can find these studies and its results in the Categorization chapter.

Once the liableness of the sources was verified we moved forward to our study. We started doing some simple linear correlation using the Spearman's coefficient. This coefficient measures the amount of correlation between two time series based on the evolution over time. It is a simple correlation analysis. If one of the variables increases over time and the other variables behaves similar in the same period of time, then the coefficient gives a value higher than 0.5, the close to 1, the more correlation exist between the two variables. Works the same way in the correlation is inverse. If one of the variables increases and the other one decreases in the same period of time, then the coefficient gets a value lesser than -0.5. This way, we should try to find the variables that have more than |0.5| with the total demand of each country. In order to simplify this we created a matrix using all the variables available for each country. The variables that we finally used for this study are:

- Gross Domestic Product
- Gross Value Added
- Direct Foreign Investment
- Interest Rates
- Employment
- Population
- GFCF

Therefore, we created a matrix with all the spearman's coefficient for all the variables and also the coefficient for the variables between each other. This

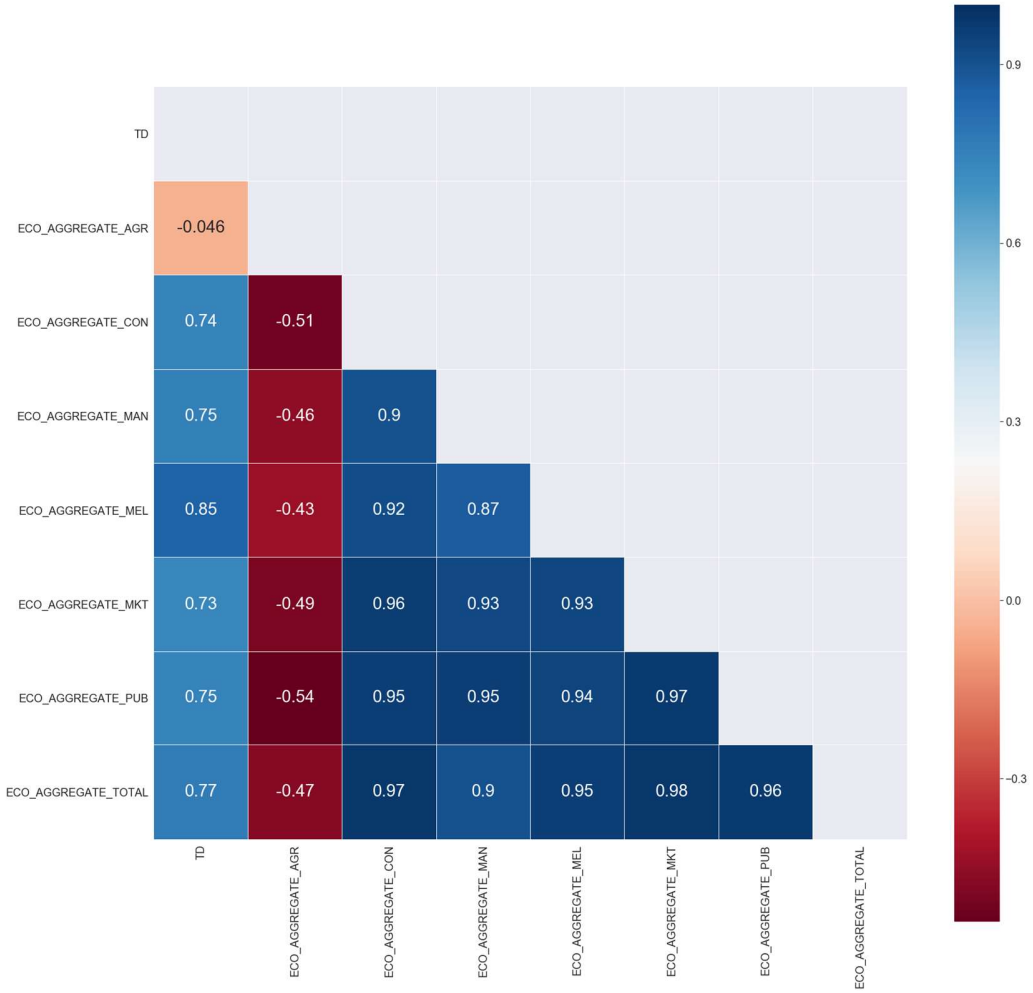
way we could also get some insights about how the country internal structure is behaving. We could see for example how the GDP is correlated with the Interest Rates. An example of this can be found in the next figure. If we want to focus on the Total demand we should read only the first column of the matrix. There is where we can find the spearman's coefficient for all the variables included in this study.



As we can observe we can find some interesting results. For example, we can see how the FDI and the Lending Rates are two most correlated variables with the Total Demand. This study is made for Indonesia. Therefore we can get a lot of information based on this results. Forthemore, we can also observe how the correlation is with the rest of the variables. The only variables that doesn give enough information that is relevant would be Interest Rates. This variables is not quite correlated with the Total Demand of commercial vehicles.

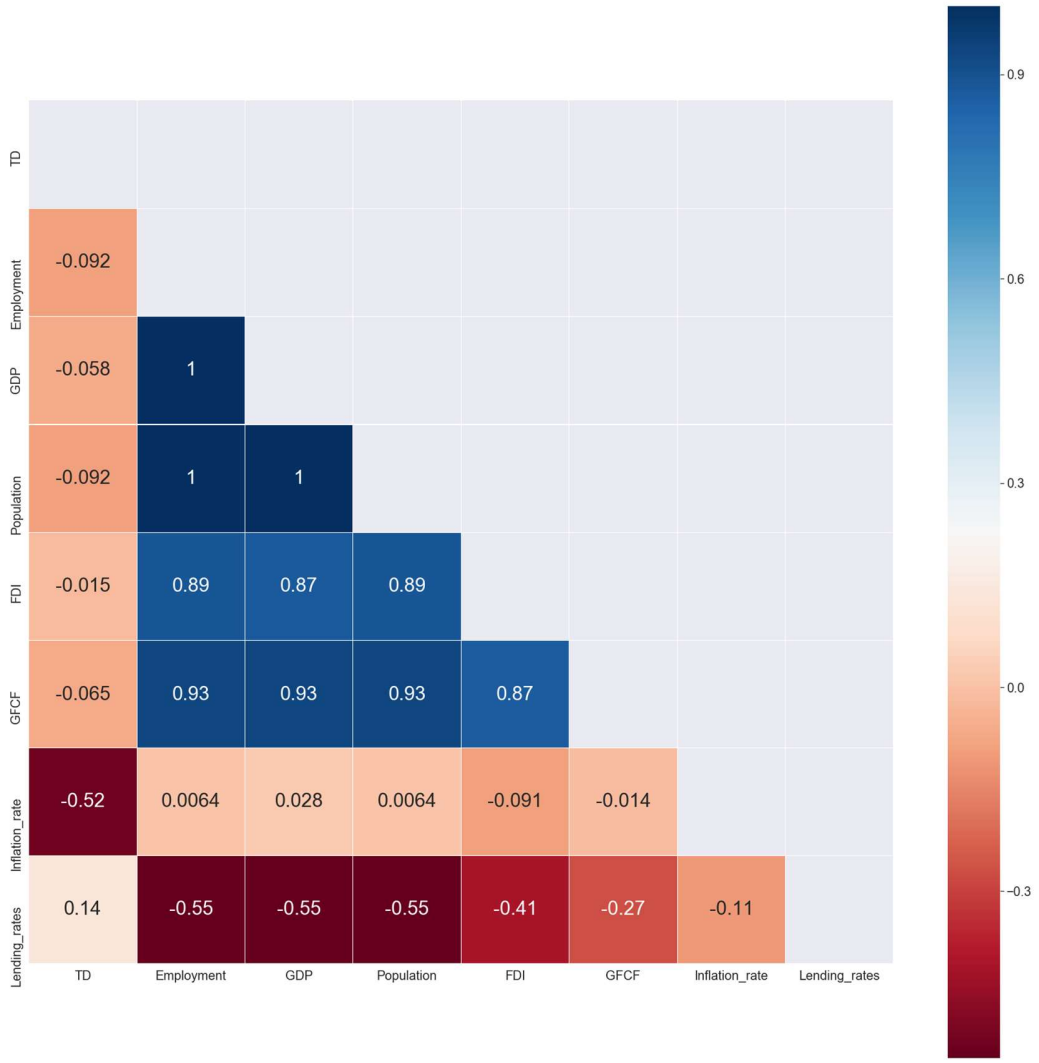
We also did some experiments with these calculations. The variable employment can be divided into the sectors of the type of employment, therefore, we can see which sector is actually causing the big correlation with

the total demand. This is useful since we want to have the results as simple as can be so when we have to look for new developments in the total demand we know where to look. In the next figure we can see the correlation matrix calculated for the same country, Indonesia, with the different sectors of employment.



We can observe that the most correlated sector is Aggregate Mel which stands for Mining and Quarring which makes a lot of sense. For the next studies we will try to implement this kind of behavior with the variables that can be divided into smaller groups. But it not secure that we would find some interesting results using these variables, since thay are only 9 variables fo explaining the evolution of a country. We have also found some not quite promising results for another countries using the same variables. We can observe, for example, how the correlation matrix is for the same variables but for Singapur:

In



this case, we can't find any variable that is enough explanatory since there is no variable that has a significant correlation with the total demand. For this case, we should try to find another variables and make the same study in order to find some variables that are explaining the behavior of the total demand.

The second part of this approach was to create a table that summarized these results. This was an idea of the Financial team. This way we could englobe all the results in one table and whenever someone needs to look to the results of one country it would be easy for them to find the most correlated variables. Inside this table we also included some technical variables like P value. Since the financial team wasn't familiar with this terminology we added to more

columns to explain this kind of variables. As we can observe in the next figure these two columns are on the right and they are called: Trust and Strong Correlation. We added some coloring in order to be easier to identify the variables that had a strong correlation. In the next figure we can observe the first version of this table.

The

	Total Demand	Variable_2	Spearmans Coeffcient	P_Value
1	TD	Employment	-0.091895	0.716862
2	TD	GDP	-0.057821	0.819729
3	TD	Population	-0.091895	0.716862
4	TD	FDI	-0.015488	0.951363
5	TD	GFCF	-0.065049	0.797613
6	TD	Inflation_rate	-0.524300	0.025508
7	TD	Lending_rates	0.136364	0.589515

information displayed on this table is the same that we can found in the first column of the correlation matrix. This table was created for the Financial team so they could get some easier insights and faster. As mentioned before we added two more columns for them to understand further the meaning behind this table.

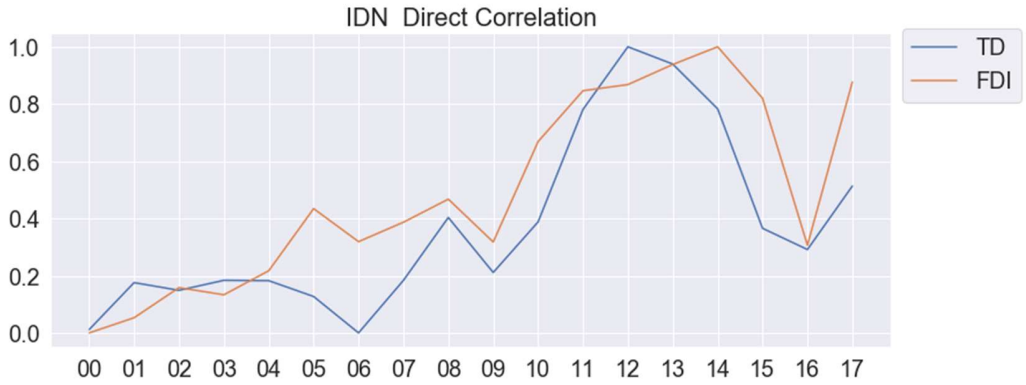
We

	Total Demand	Variable_2	Spearmans Coefficient	P_Value	Strong Correlation	Trust
1	TD	Employment	0.770898	0.000180557	False	True
2	TD	GDP	0.76677	0.000205354	False	True
3	TD	Population	0.76677	0.000205354	False	True
4	TD	FDI	0.810114	4.59455e-05	True	True
5	TD	GFCF	0.76677	0.000205354	False	True
6	TD	Inflation_rate	-0.459221	0.0552248	False	False
7	TD	Lending_rates	-0.811488	4.35531e-05	False	True

added this two columns: Strong correlation and Trust. These two columns represent the Spearmans coefficient column and the P_value column. To these columns we finally added some coloring so the result was really easy to read and they would only need to focus in the green cells. An example of this table can be found in the next figure.

Total Demand	Variable_2	Coefficient	P_Value	Correlation	
TD	ECO_AGGREGATE_AGR	-0.0464396	0.854814	False	False
TD	ECO_AGGREGATE_CON	0.744066	0.000399211	False	True
TD	ECO_AGGREGATE_MAN	0.754386	0.000297661	False	True
TD	ECO_AGGREGATE_MEL	0.853457	6.68154e-06	True	True
TD	ECO_AGGREGATE_MKT	0.733746	0.000528393	False	True
TD	ECO_AGGREGATE_PUB	0.74613	0.000376859	False	True
TD	ECO_AGGREGATE_TOTAL	0.770898	0.000180557	False	True
TD	Employment	0.770898	0.000180557	False	True
TD	GDP	0.76677	0.000205354	False	True
TD	Population	0.76677	0.000205354	False	True
TD	FDI	0.810114	4.59455e-05	True	True
TD	GFCF	0.76677	0.000205354	False	True
TD	Inflation_rate	-0.459221	0.0552248	False	False
TD	Lending_rates	-0.811488	4.35531e-05	True	True
TD	ECO_DETAILS_A	-0.0464396	0.854814	False	False
TD	ECO_DETAILS_B	0.907121	2.0737e-07	True	True
TD	ECO_DETAILS_C	0.754386	0.000297661	False	True
TD	ECO_DETAILS_DE	0.651187	0.00341978	False	True
TD	ECO_DETAILS_F	0.744066	0.000399211	False	True
TD	ECO_DETAILS_G	0.667699	0.00246239	False	True
TD	ECO_DETAILS_HJ	0.159959	0.526061	False	False
TD	ECO_DETAILS_I	0.628483	0.0052147	False	False
TD	ECO_DETAILS_K	0.717234	0.000807071	False	True
TD	ECO_DETAILS_LMN	0.78741	0.00010501	False	True
TD	ECO_DETAILS_O	0.624355	0.00561127	False	False
TD	ECO_DETAILS_P	0.729618	0.000589052	False	True
TD	ECO_DETAILS_Q	0.723426	0.000690902	False	True
TD	ECO_DETAILS_RSTU	0.706914	0.00103676	False	True
TD	ECO_DETAILS_TOTAL	0.770898	0.000180557	False	True

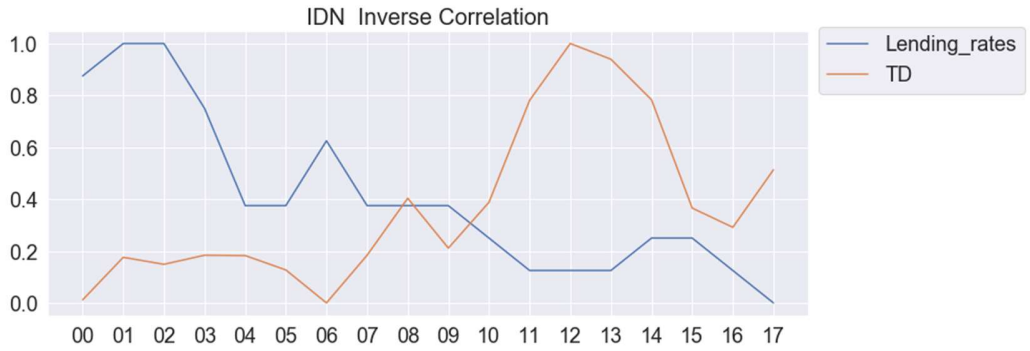
The last part of this approach was to create two plots with the results of the correlation matrix. In order to see and validate that the results were correct. We created two plots, one for the direct correlation and another one for the inverse correlation. This way we can clearly see how the evolution over time of the total demand of commercial vehicles is correlated with the different variables. An example of the direct correlation can be found in the next figure.



We can clearly see how the two variables are correlated over time. What we don't know is which one is causing which. But we can get a clear idea of how these to variables behave over the time that we studied. We can also get some insights about the country behaviour since the FDI is definitely influenced by other economic variables. We can for example see how in 05 the FDI increased

too much that the total demand decrease. This would be an interesting study for next approaches.

In the next figure we can see a similar plot regarding the inverse correlation.



Here we can clearly see how the inverse correlation works. When the lending rates go down the Total Demand increases drastically. Similar happens in the Lending rates increase. These two plots are also part of the final deliverable that we handled to the Financial team. We would get more information about the final result of the project in the Results and Conclusion chapter.

f. Singular Spectrum Analysis

Once we had cluster the data and therefore made it more accessible, we started with new methods. This time trying to explain the evolution of only one time series. For this, we started a decomposition approach. The idea is to decompose the time series into Trend, Seasonality and Noise. Once we have done this, we can try another approaches combining both approaches. For example, we could try to cluster the time series based only on its trend or on its seasonality. This way we could find which countries are affected by the same events. We didn't follow this approach finally because we used another approach for the events correlation study. Nevertheless, we achieve a good performance in this method. It can be also used for a whole cluster. Therefore we can see if some countries are clustered in the same cluster by kmeans because of them trend of because of them seasonality.

A continuation, we would explain the theory behind the decomposition method used for the decomposition. We used a method called Singular

Spectrum Analysis (SSA). Which is defined as a nonparametric and adaptive spectral decomposition of a time series[6]. The SSA method decompose a time series in different components that, after some modifications, can be classified as Trend, Seasonality/ Periodicity and Noise.

The theory behind this method is complicated. We would try to explain each point carefully[7]:

The first step is decomposing the time series in subseries. Given a window length, we create as much as subseries as we can:

$$
\begin{aligned}
X_0 &= (f_0,\ f_1,\ f_2, \ldots,\ f_{L-1})^{\mathrm{T}} \\
X_1 &= (f_1,\ f_2,\ f_3, \ldots,\ f_L)^{\mathrm{T}} \\
X_2 &= (f_2,\ f_3,\ f_4, \ldots,\ f_{L+1})^{\mathrm{T}} \\
X_3 &= (f_3,\ f_4,\ f_5, \ldots,\ f_{L+2})^{\mathrm{T}} \\
&\vdots \\
X_{N-L} &= (f_{N-L},\ f_{N-L+1},\ f_{N-L+2},\ \ldots,\ f_{N-1})^{\mathrm{T}}.
\end{aligned}
$$

Figure 13 - Window Subseries

With this windows we create what is called a Trajectory Matrix. This matrix is a Hankel matrix, given that all of the elements in the diagonal *i+j = const* are equal. Once we have the Trayectory Matrix we have to use the SVD technique (Singular Value Decomposition) in order to find the orthonormal matrix inside the Trajectory Matrix. A graphic example can be seen in the next Figure.

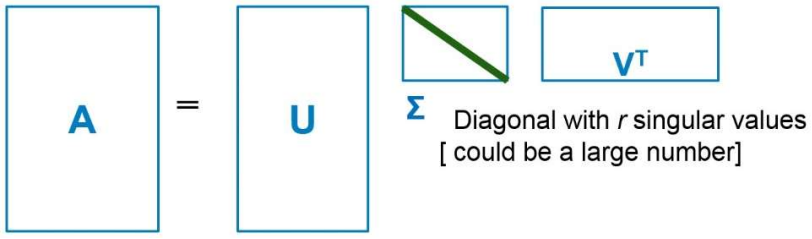


Figure 14 - SVD of Trajectory Matrix

As some authors have declared[8], there are some similarities between this method and the Fourier analysis of Time series.

Once we have this SVD we can divide the X matrix in sub X_i matrix. With this methodology, we will have to choose the correlation between the X_i using a visual inspection. Once we have all the sub matrix we see how they are distributed and we can group them. But there is another method.

Once we have divided our time series in sub series like F_i we can create a matrix called W_{corr} that calculates the correlation between this sub series. This way we can get the relation that each time series has with each other.

After this step we get the W_{corr} , which can be found in Figure 13.

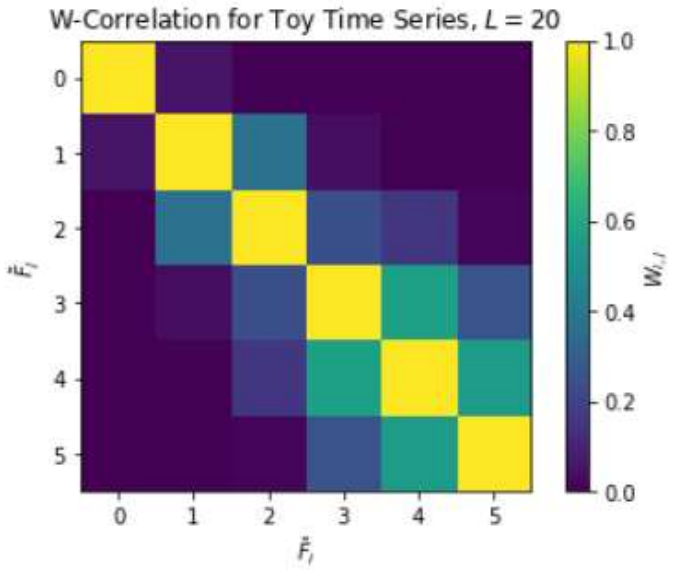


Figure 15 - Trajectory Matrix

As we can see we can cluster the 5 mini time series create by the SSA method in different ways. The most normal one would be to cluster 0, [1,2], [3,4,5]. If we decide to create this clustering we would get something like the Figure 14.

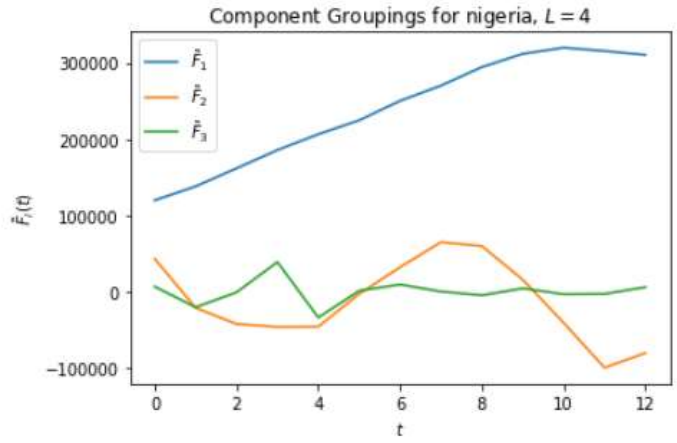


Figure 16 - SSA Clustering

We can observe that the trend component is well defined with this clustering of the mini time series created by the SSA method but the Noise and the seasonality are not quite differentiated. This is the reason why we should decide another type of clustering of the mini time series until we can perfectly distinguish the 3 principal components that form a time series. The final result of this approach can be seen in the Figure 15.

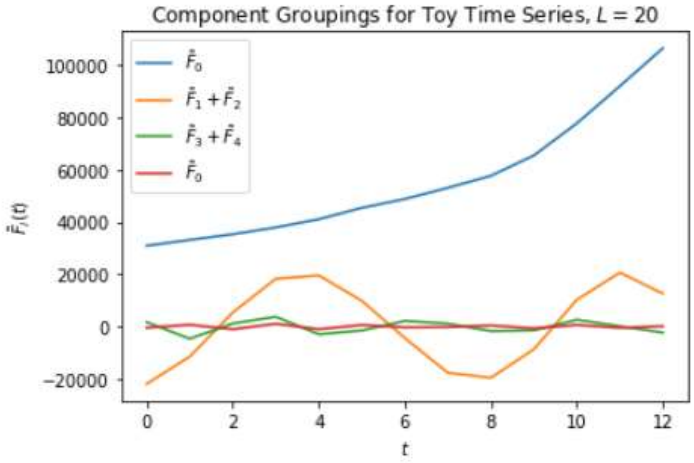


Figure 17 - SSA Clustering Components

Here we can perfectly observe the three components that are present in every non-stationary time series. This is a really good solution for this approach, now we can automatized this process so we can do the same for all the time series of a cluster. That way we can do the approaches explained before. If we choose a simple cluster with only three time series, in order to be easy to see all the components of all the time series inside of it. We can observe the solution obtained in next figure.

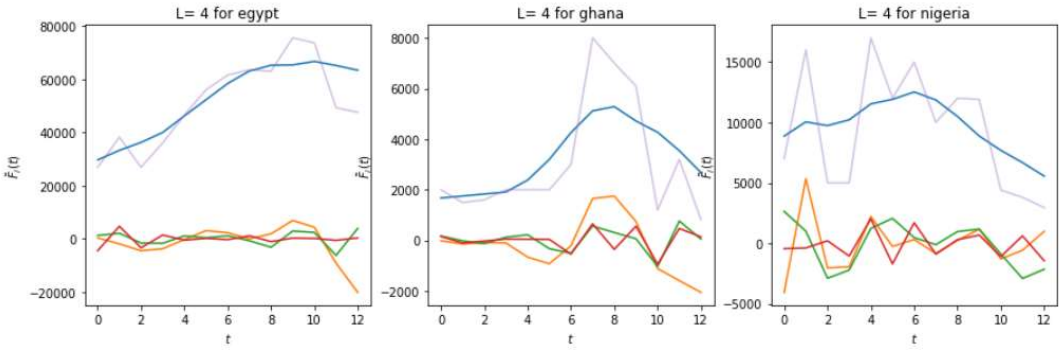


Figure 18 - SSA for whole cluster

g. Dynamic Time Warping

In the next two chapter we would study the two methods that were the most complicated part of this project. We dedicated quite a lot of time for its deployment and they have a lot of potential for futures development. Nevertheless, the business department had other needs and the project had to continue in another direction. We would summary they achievements of this project on the next chapters. Our goal here is to explain the implementation of this two methods in order to create this line of work for future developments. The math behind this two methods will be explained using figures and functions in order to simplify the compression of these new approaches to treat time series.

In this first chapter we would study a new approach to cluster time series, using an optimal match technique in complement with a holistic matrix as we would see a continuation. In the next chapter, we would focus on the utilization of this cluster created by the DTW method. The solution of both projects would be found at the end of the DBA chapter.

The DTW is basically another method to measure the distance between two time series. Normally the distance between to time series is calculated using Euclidian distance or Manhattan distance. This is normally a linear method that compares the *i-th* point of one of the time series with the same *i-th* of the other time series[9]. Therefore, we can get a value that gives us the distance between the two time series but this is not perfectly reliable since two time series can be similar but being retarded one to the other. For example, in the Figure 19 we can observe the difference between using the Euclidian method and the DTW method for calculating the distance between two times series that are slightly different.

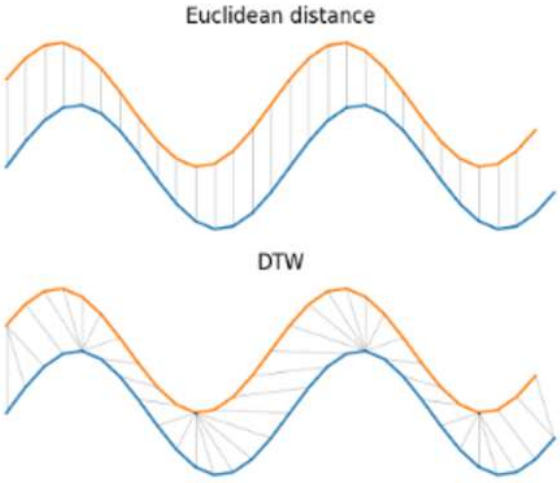


Figure 19 - Euclidian DTW difference

Therefore, the Euclidian distance won't the best method to calculate distance between some time series. If for example, we want to find if some time series is equal to the same time series but slightly fast, like the same person talking slowly and fast, the Euclidian distance would find some great differences between the fast and the slowly one. This is way, the DTW was first intended to be a method to find similarities between speeches of different people. It can be useful for voice recognition since it takes another method to measure the similarity between time series and therefore we could categorize the same voice is different speed for the same user. In Figure 20 we can observe two examples were the Euclidian wouldn't be ideal[10].

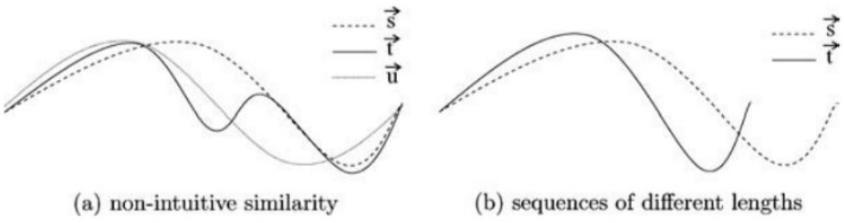


Figure 20 - Ineffectiveness of Euclidian distance

The DTW has a couple of steps that should be follow if we want to achieve a good solution[11]:

- Divide both series in the same n points.
- Calculated the distance between each point of the first time series with all the points of the second one.
- Repeat the previous point with second time series as the base.
- Add all the minimum distance for each point.

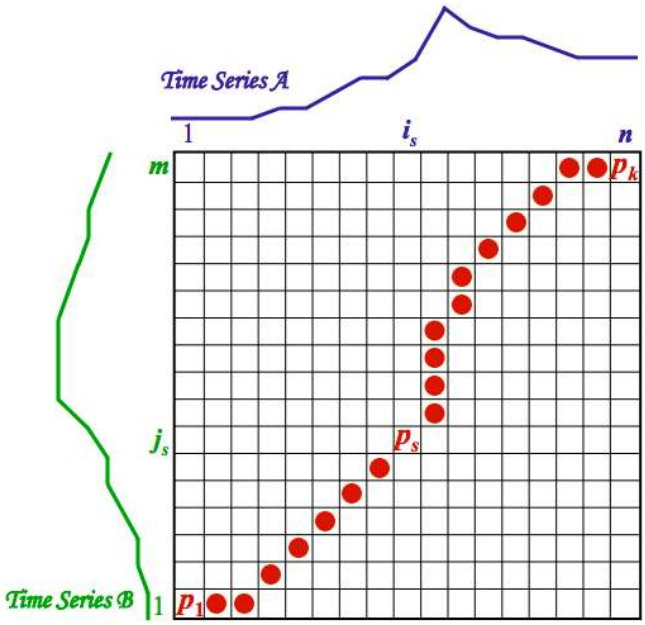


Figure 21 - DTW between two time series

Using these simple rules we find the path that has the minimal distance between the two time series. This is going to be really different that the Euclidian distance if we have some speed or time difference between the time series.

Although, there are some restraints we have to take in account while using this method in order to optimize its output. This restriction have, mostly, one objective: reduce the search space. This means, reducing the number of points where the path of minimum distance can be stored, which is achieved by defining the next four restrictions:

- Monotonicity:
  The alignment path can't go back in time. This means that we will always compare the i-th iteration of one of the time series with the i-th + j. The reason after this is that guarantees that one feature won't be align with two features of the other time series:

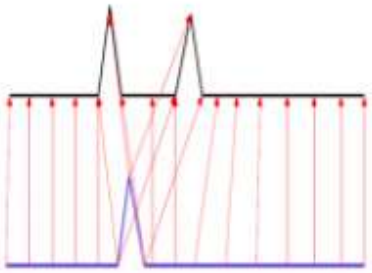


Figure 22 – Monotonicity

- Continuity:
  The path can't jump in the time. Every i-th of the time series has to be aligned with another point. This assures that no feature is missed by the algorithm:

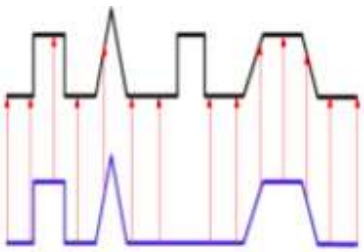


Figure 23 - Continuity

- Boundary Conditions:
  The path starts a point (1, 1) at finishes at point (n, m). This guarantees that all the time series is assessed by the method.

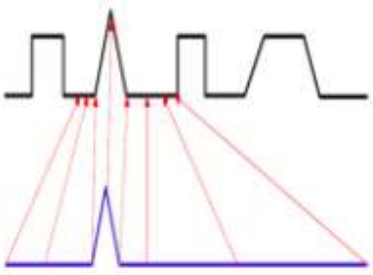


Figure 24 - Boundary Conditions

- Warping window:
  We create a window length that guarantees that the path doesn't wonder to far from the diagonal, given that the optimal path won't be too far from the diagonal.

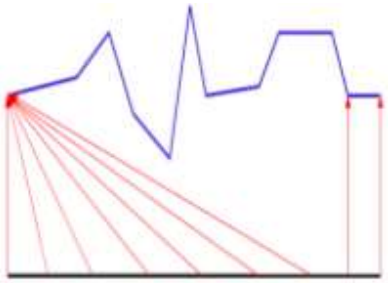


Figure 25 - Warping Window

- Slope Constraint:
  The alignment path can't be perfectly horizontal nor perfectly vertical for more than 3-4 points. This assures that short parts of the first sequence are not align with long sequences of the second time series.

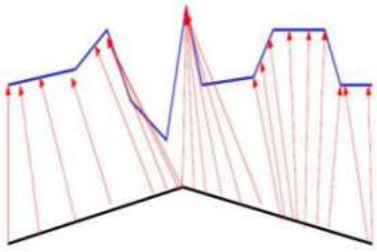


Figure 26 - Slope Constraint

Apart from this constraints we have another one regarding the selection of the weights used by the algorithm in order to time-normalize the distance between the two time series. The only constrains that we have to take in account here is that the summation of all the weights has to be equal to value C that verifies:

$$D(\mathcal{A}, \mathcal{B}) = \frac{1}{C} \min_{P} \left[\sum_{s=1}^{k} d(p_s) \cdot w_s \right]$$

Figure 27 - Weights constraint

A last image of the method finished can be seen in Figure 28, were we can see how all the points are looking for the closer path to connect all the points. It is important also to remark the direction of all the points. We can observe that once all the calculations are done. We begin in the last point and finish the minimal path backwards.

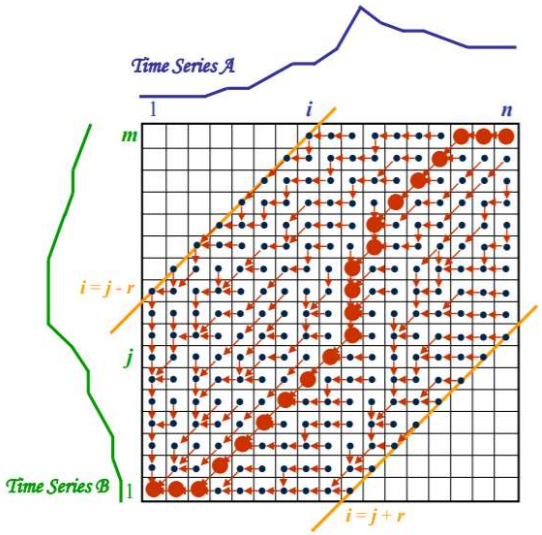


Figure 28 - DTW method

Once we have explained the theory behind this method we can show the results obtained for this project. The idea we have with this method is use this DTW methodology in close relation with clustering. Therefore, when before we had seen that we were using the Kmeans method to cluster the time series in different cluster that had similar time series with the only objective of being able to have a better visibility of some of the countries. Here we want to achieve something similar but using the distance calculated by the DTW method as the main variable to cluster the time series. We have found there is a package inside the tslearn library of python that does something similar. It's called *GlobalAlignmentKernelKMeans.* Even if the name of the method has the

kmean on it, it has some input where you can decide the metrics you want to use to cluster the time series, this metric is called sigma_gak in this method. We would study this in more deep when we arrive to the chapter of resources.

We found some interesting results but it was really difficult to minimize the importance of the volume of the time series. We tried standardized the data, but since we lost too much information when we were standardizing the data from a cluster we ended up doing this method for all the time series of a cluster with the actual values. This way, the kmeans method would minimize the difference between the values of the time series and then the DTW would try to obtain the distance of the time series that are already cluster by the kmeans. In the Figure 28 we can find to clusters created by the method mentioned before.

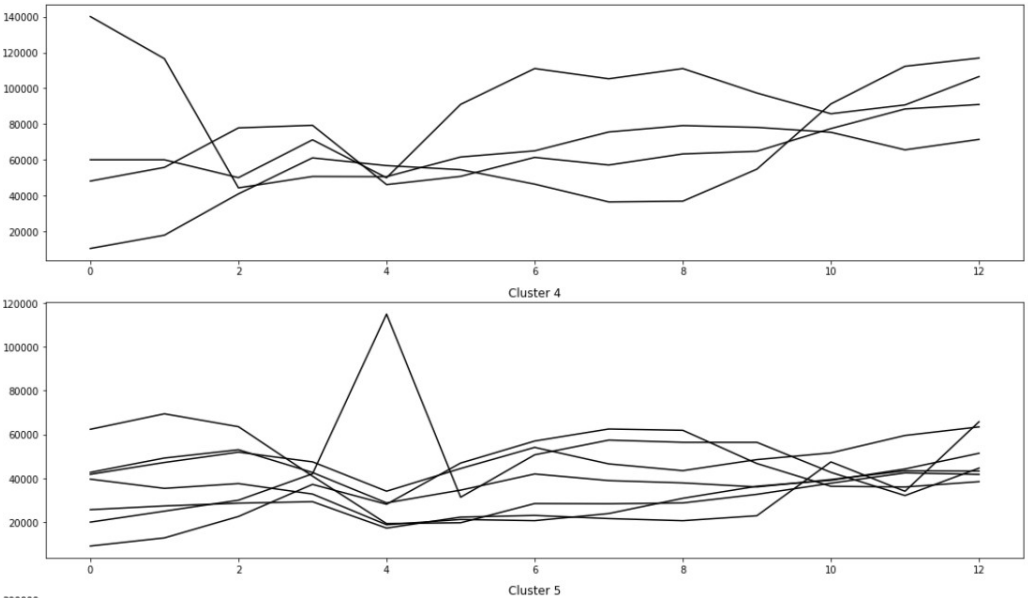


Figure 29 - DTW Clustering

This two examples are really good because we can observe how the method is not perfect because, as we can see in the first one there are differences between the time series that are being cluster but we have a similar evolution of tem. Which was the point of using this method. In future developments, the idea would be to optimize the calculations created by this library and try to cluster only the time series that have a similar evolution, disregarding the ones that have a small difference but different evolution since what were are after is the time series that a similar evolution over time.

After this clustering, we can now start with the next method used in this project. Which uses this clustering done with the DTW to create another *cluster* but with a different output.

h. DBA

The DBA method, Dtw Barycenter Averaging, consists in iteratively refining an initially average sequence in order to minimize its distance to the averaged sequence[12]. This method consists in a iterative procedure to summarized the data into a medoid. A mediod of a time series will be the point which minimizes the DTW distance with all of the time series taking in account in the DBA method[13].

A more simple definition would be, a method that tries to minimize the distance between the time series solution and all the time series that are inside the cluster it is trying to average. So for example, in Figure 30, we can see the average time series between two datasets. The idea will be to compute the barycenter between each of the segments that compose the two time series.

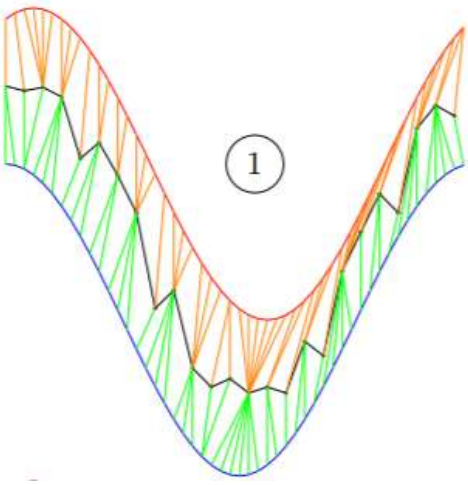


Figure 30 - DBA Method

It is important to note that one coordinate of one of the sequences can contribute to the calculation of the barycenter of other coordinates, as we can see in figure 30. After this calculation is made, the algorithm starts with the optimization of the location of this barycenter, for each iteration DBA works in two steps:

- Compute DTW between the sequences.
- Update the coordinates of the average time series as the barycenter of the coordinates of the base time series.

After iterations, the average time series will be smoothed, given that the error in the stimation minimizes everytime we create a new set of barycenters for the two time series. In the next Figure, we can see the result of this DBA method after 4 iterations.

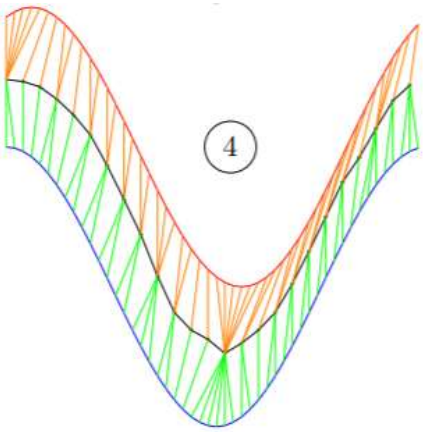
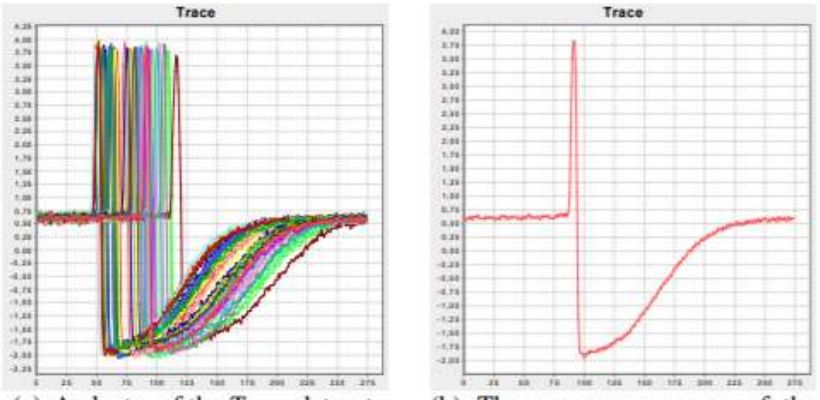


Figure 31 - 4th Iteration DBA

Since the update of the DBA method is made for only two time series each time, and each update is independent from the next ones. The DBA stands as a very good and reliable method for averaging a large amount of time series at the same time. We will find the barycenter between two time series and continue for the next one. Since the iterations work independent, the solution would be an average of all the time series taken in account in the cluster.In figure 32 we can se and example of this methodology.



(a) A cluster of the *Trace* dataset. (b) The average sequence of the cluster.

Figure 32- DBA for cluster

The reason after choosing this method is the performance that it has. It has been found that the DBA has a better indicators than other averaging method for time series. There is some work done by Saeid Soheily-Khah, for example, that demonstrates the outstanding performance of this method in comparison with methods like NLAAF, PSA or CWRT. We can see in Table 13 a table comparing the values of the IRR (Inertia Reduction Rate) which is a measure of how representative the extracted centroid from the method would be.

DATASET	NLAAF	PSA	CWRT	DBA
CBF	8.3%	12.3%	-61.3%	**32.1%**
CC	9.8%	28.6%	6.8%	**34.2%**
Digits	26.1%	79.5%	77.6%	**82.2%**
CHAR. TRAJ.	67.1%	87.7%	85.2%	**90.6%**
BME	34.9%	43.1%	-11.8%	**59.4%**
UMD	25.6%	**51.1%**	-56.2%	48.8%
SPIRAL	59.8%	64.4%	64.2%	**65.8%**
NOISED SPIRAL	61.4%	**66.3%**	9.3%	9.8%
CONSSEASON	**84.1%**	70.5%	4.6%	21.4%

Table 13 - IRR comparison for DBA

Therefore, once we have explained the theory behind the method and explained also why we have decided to use this methodology we can now go on and study the solutions we obtained with this approach in this project. It is important to remember that we have created cluster based in the distance calculated with the DTW. This method would use the clusters created by the DTW method to calculate its average also using DTW. Therefore, there are two different uses of DTW. The first is in relation with the methodology to cluster the total demand of the countries and the second one is to calculate the barycenter of the time series inside the clusters created by the method mentioned previously.

We can observe then, the DBA method applied to the cluster shown in the Figure 29.

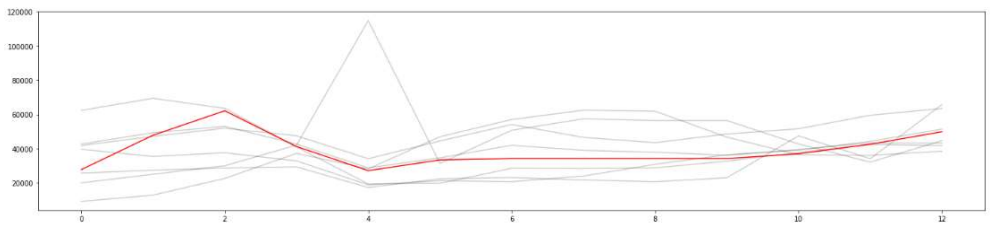


Figure 33 - DBA for cluster

We can see that the solution showed by the DBA method is quite interesting. Given that the DTW wasn't functioning perfectly since it cluster that time series

with that peak at 4, by using the DBA method we can smooth the error since the majority of the time series inside the cluster have a similar behaviour, the impact of the peak is reduce by the effect of all the other time series. Therefore, it is a good approach to use this two method together, since the DTW clustering would have some errors regarding the evolution of all the time series clustered but the DBA method would smooth this errors and obtained a perfectly good time series that summarizes more or less the global behaviour of the cluster. We can observe in Figure 34 another example of this DBA method.

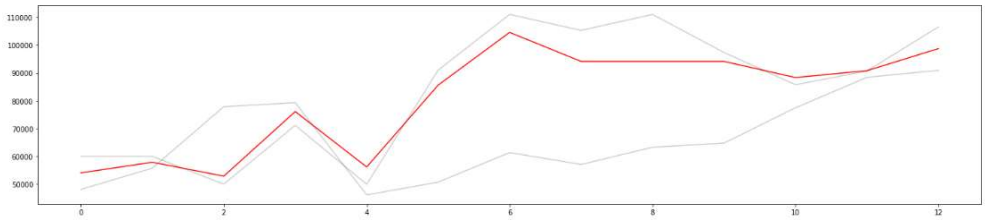


Figure 34 - DBA for two time series

Here we can observe that the performance of the DBA is far from been perfect, since the upper time series is having a much more impact in the averaging time series that the lower one. This can be also explained since the trend of the lower time series after the point 4 is positive. Therefore, the algorithm can compute this as a close barycenter to the upper one, we can also observe how in the point 10 the average time series goes upper than both of the time series. This can be explained by the idea we have explained before that is that one barycenter can be affected by some precedent events. What is most likely happening there is that the trend in the lower time series is influencing in the averaging method in a way that computes the averaging time series with a higher value of the actual two time series taken in account.

The last approach we did with this method was to customize the code so it could be applied to every cluster inside a cluster. If, for example, we would have a kmeans cluster that could be divide in three cluster by a DTW methodology, then we would use the DBA for each one of those three cluster and therefore minimizing the dimension from 7 or 8 countries to 3.

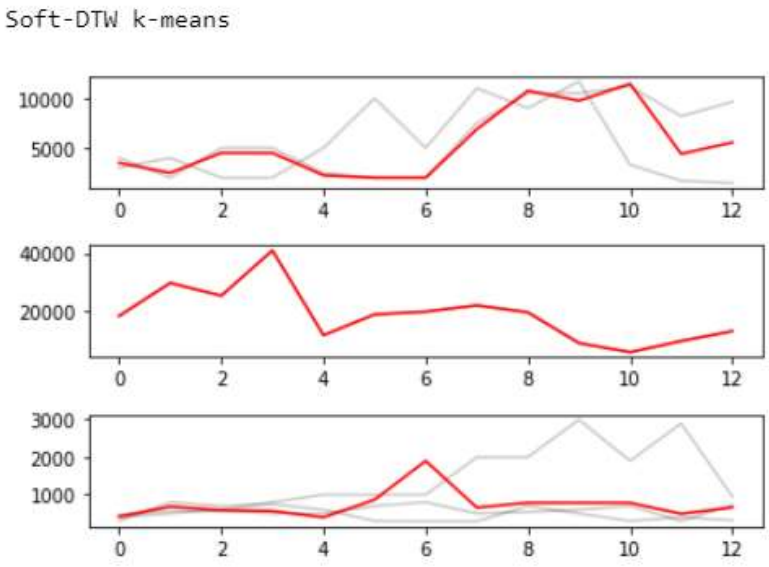


Figure 35 - DBA for cluster of cluster

This is also in the same direction of the PCA method, the idea at this point was to reduce the datasets where we were working on so we could try to make some kind of forecast for only this reduced dimensions and then go back in our footsteps and apply the forecast values to the initial values in our datasets.

i. Monthly Data Approach

After the study of all this methods, we tried to use another type of data based on the problem that we didn't had enough values for achieving some interesting results with some method. So far we were working with 25 years, which was okay for most of the methodologies but also was a little bit concise with the results. Given that the next approach was going to be focus on creating a forecast. The business department and the big data department arranged that another set of data would be needed for this approach. Because of this reason we tried to find some kind a documents that were similar to the previous ones, so we could adapt the code easily, but with bigger dimension. We founded that, for most of the variables that we were taking in account, we could find monthly data so we started to study all the methods that had been presented here with these new dataset with monthly data.

The first challenge we had with this data was the merging of two datasets. In order to have monthly data of the total demand since 97 we had to create a simple script to simplify and clean both datasets and then merge them. We

had to do this for three years: 1997, 1998 and 1999 so they could be displayed together with the values from year 2000.

	Country	JAN_97	FEB_97	MAR_97	APR_97	MAY_97	JUN_97	JUL_97	AUG_97	SEP_97	...	Apr_17	May_17	Jun_17	Jul_17
0	Austria	1968	1606	1902	2404	1836	2234	2058	1649	1831	...	3211.0	3674.0	3787.0	3406.0
1	Belgium	4627	4544	4370	4934	3868	4226	3247	2671	2980	...	6712.0	7493.0	8297.0	5697.0
2	Denmark	2316	2104	2217	2560	4490	2015	1850	1906	2370	...	2816.0	3363.0	3715.0	2119.0
3	Finland	1374	839	950	1298	995	1173	865	1009	1166	...	1307.0	1480.0	1428.0	904.0
4	France	29945	25962	27240	31359	23608	26042	22287	19925	24516	...	36201.0	36738.0	46221.0	32172.0

5 rows × 274 columns

Table 14 - Monthly Data

As we can see we would have 274 values from the year 1997 to year 2017. This values will be for Western Europe, since the reliable and trustworthy dataset could only be found for this type on country that have had a stability over the years so this data could have been recorded. For Asia, since is the market we are most interesting in, we won't have enough data to create this datasets. If ever the approaches taken in this chapter ended up being much more accurate than the previous one with less amount of data, we would have to study the possibility of looking for this kind of information for the countries that we are interested on, or start some approach considering dividing the actual data into smaller pieces to fool the algorithm about the volume of values available.

We continue with this dataset for now. In order to cluster the data using kmeans we have to standardize it. We will use the same method described before, we will use the sklearn library and the Standard Scaler method. This way, all the values inside the dataset would be inside a range between 0 and 1. This is useful to cluster the data without taking in consideration the huge difference between the contries. We can see a plot of one of the clusters in Figure 36.

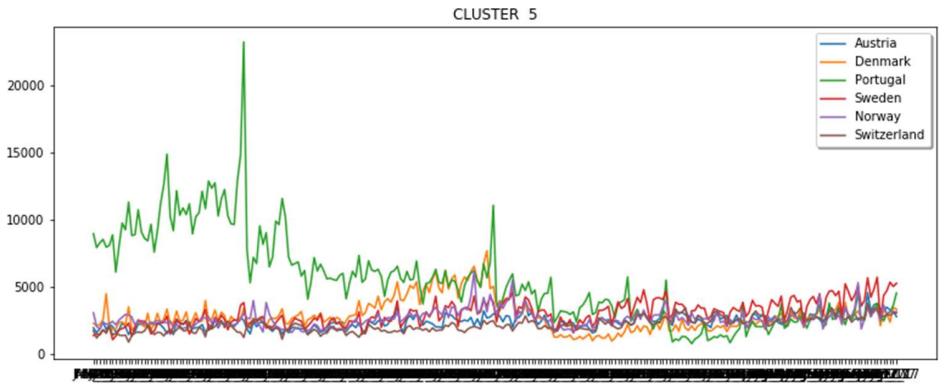


Figure 36 - Monthly Kmeans Clustering

We can observe that this data is going to be more difficult to understand since the evolution between years is more complex that the evolution seen before. This is why we wanted to use this type of data so the method that we have studied could give some insights about the actual evolution of the Total Demand of Commercial Vehicles per year and even in a five year rate. We can see for example, the evolution of the Total Demand for Spain to see the evolution of it.

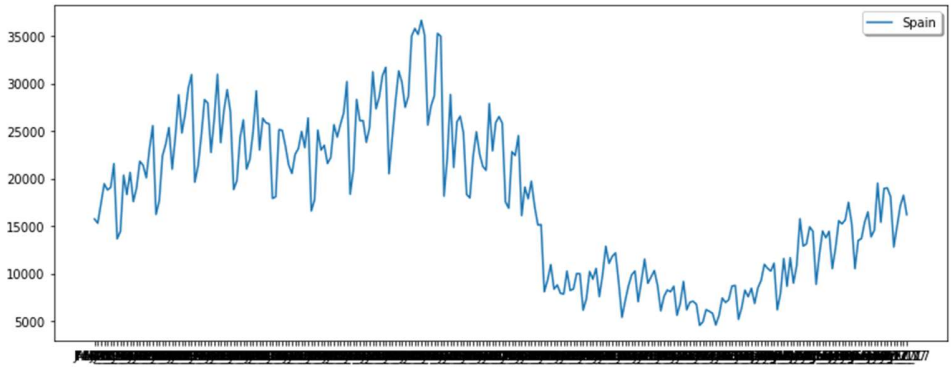


Figure 37 - Spain monthly Evolution

Since the data is so large, it was a difficult task to be able to display the labels for every value. We tried creation a rotation of 45 degrees so we could at least see the year we are in, but since the idea was to know in which month the change was taking place. We couldn't find any proper solution to this problem. We can see for example that Spain has a really big decrease in the Total Demand, we suggest this is around the year 2009 when the crisis was taking place. This would be make a lot of sense. Also, we ca see how the seasonality is really strong, we can see how the fluctuations are very similar from one year to another. WE would be able to see this with the next method utilized with this dataset.

We will use the SSA analysis again to be able to decompose the time series in sub series so we can see which behaviour is going to be predictable and which one is going to be random and we have to be careful and study those kind of behaviour to try to find some kind of explanation for them, decreasing then the amount of randomness or inexplicable data. Which is the main purpose of this project, to understand more the drivers of the Total Demand of the Commercial Vehicles. A first iteration to the SSA analysis will give us a decomposition like the one found in Figure 38.

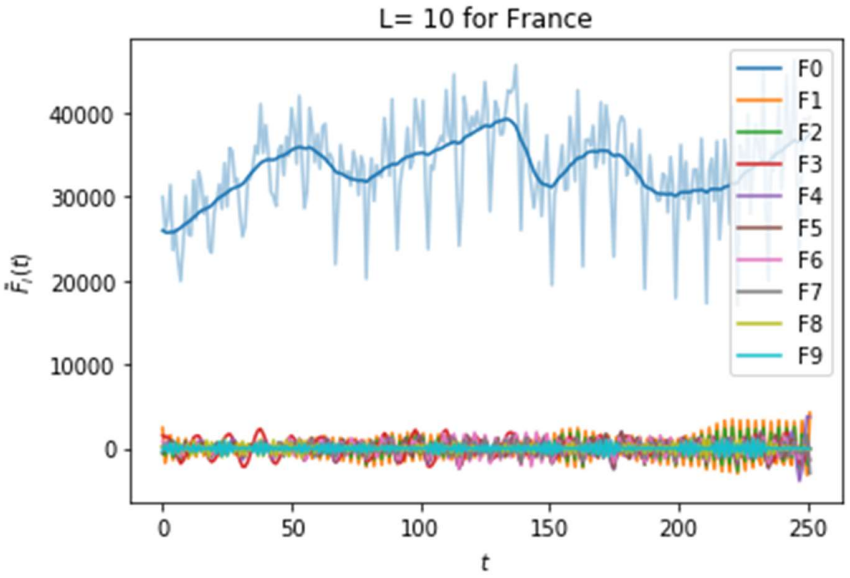


Figure 38 - SSA for Monthly Data

We can see that the trend is perfectly described with this decomposition. We can't observe however the differentiation between the noise and the seasonality in the rest of the sub time series created by the SSA method. We will try to study the Correlation Matrix in order to find which sub series we should merge in order to find a more readable plot.

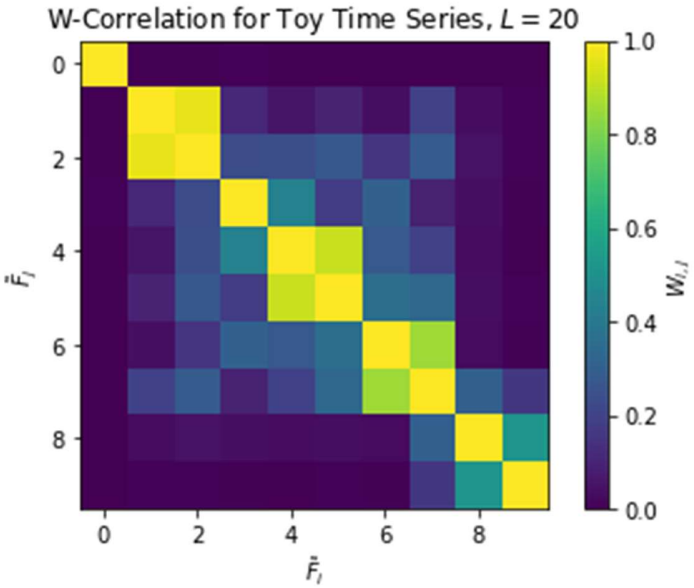


Figure 39 - Correlation Matrix for Monthly Data

With this matrix we can merge different spectrums signals in order to create three different components of the main time series that can resemble as trend seasonality and noise. They would never be the same as this components since the would be also a time series but the idea is to have 3 time series that explain the behaviour of the main time series as good as possible. After a couple of iterations we obtain the grouping shown in Figure 40, which explains the evolution of the Total Demand.

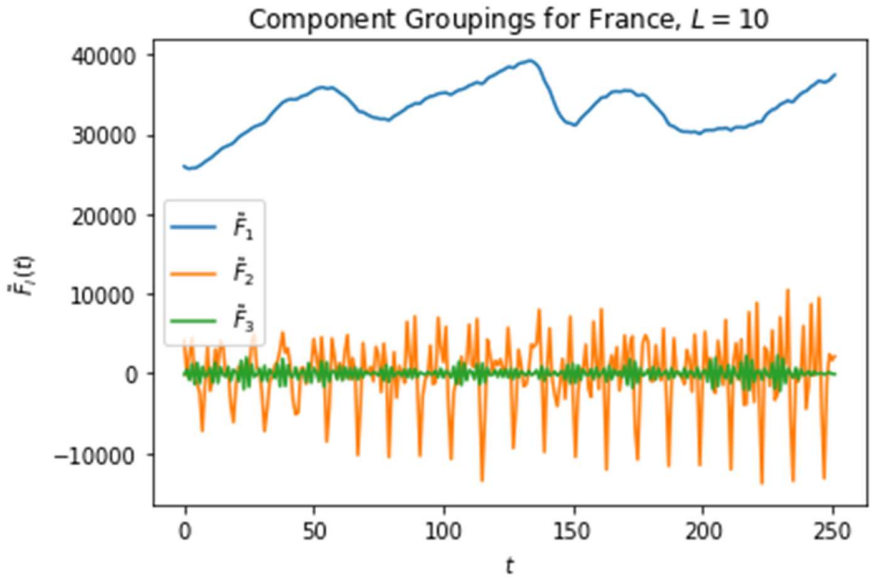


Figure 40 - Component Grouping for Monthly Data

Once we had the time series decomposed in the tree basic components, we can move to next step of our project which was creating a Forecast. This was thought as a solution that could tell how the Total Demand was going to behave in the next few years and therefore, from that information, extract if the correlation analysis we made was reliable or not. If for example, we based the evolution of our Arima model in the variables that we have found during the correlation study and then we find that the forecast is not accurate enough we can inherit from that, that our correlation study does not have a good performance. Therefore, even if the forecast is not going to be the main aspect of this project ultimately, it would be useful to evaluate the rest of the method taken in account during the project.

j. Forecast

The theory after the forecast is simple but it can turn into really complicated methods. Here we will primarily study the Arima Model, since it is the most

robust once and it can be also use as a multivariate model, which could be really interesting. For example, create a forecast of the Total Demand using a multivariate analysis and taking in account the most correlated variables. This approach would be part of this chapter.

The ARIMA model stands for: Autoregressive Moving Integrated Moving Average. Actually, it is a model that agglutinates three different models. We will try to summarize each one of this a continuation in order to explain the Arima model later on this chapter.

$$Z_t = C + \sum_{i=1}^{p} \phi_i Z_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Figure 41 - ARIMA Equation

In Figure 41 we can observe the different parts of the Arima Equation. The C is the constant part of the model it is achieve after some computation that is needed to transform the initial values of the time series into a stationary time series, condition that has to be met in order to apply the model. The ε_t is the error of the element t. The first summation [1]:

$$\sum_{i=0}^{p} \emptyset_i Z_{t-i} \quad [1]$$

Represents the AR, Autoregressive Model, terms, which are the lagged values of the time series. This means that we compute the precedent values of the t-th iteration of the time series to calculate its prediction. The second summation [2]:

$$\sum_{i=0}^{p} \theta_i \varepsilon_{t-j} \quad [2]$$

Represents the MA, Moving Average, terms, which represents the lagged errors. Which means the error of the predictions made by the model for previous values. This way, we take in account the previous values of the time series as well as the errors made by the model in order to calculate the predictions for the previous values.

We will now study this two methods in order to understand how the Arima model works.

- Autoregressive Model:

The AM model predicts the future values of a time series based on the previous values of the time series, plus a stochastic value which give the model the degrees of randomness needed to address a stochastic time series. The model is basically a linear regression of the current time series with the precedent values of the same signal[14].

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots\ldots \phi_p Y_{t-p} + \varepsilon_t$$

Figure 42 - Autoregressive Model Equation

- Moving Average:

  The MA method is a calculation of the average of previously calculated segment of a particular time series. The time series that is being analyze gets divide in some sub series and we compute the average of all the values of each sub series creating a vector of averages. It has different possibilities, such as: Simple moving average, weighted moving average and exponential moving average. All this different approaches differ in the way the averages are stored inside the vector of averages.

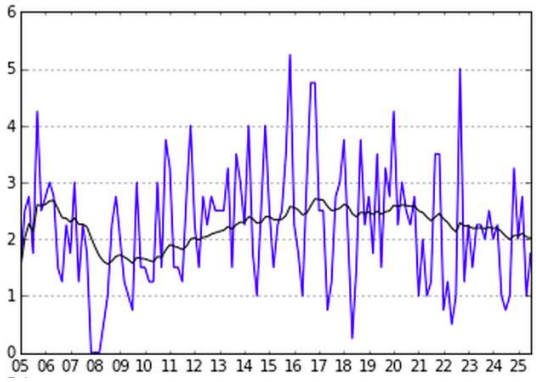


Figure 43 - MA example

In the example shown in Figure 43 we can observe how the MA method simplifies the time series by calculating the average of the previous points for each year. This gives a really good and visual idea of how was the behaviour of the variable during the whole time slot. It would also filter the noise and the fluctuations of the time series in order to achieve more clarification when studying time series.

Once we have understood both of the method that are used inside the Arima Model we can study its behaviour and implementation. There is one last part

of the method that we have to explain before starting to assess the actual behaviour of the Arima model: the integration.

- Integration:

  The Arima model only differences itself from the Arma model in this integration. This integration part is needed when the time series that we want to study doesn't have a stationary behaviour. This stationary behaviour is important since it is one of the requirements of both AR and MA models to function. Therefore, we have to implement a differencing step in order to achieve this stationary behaviour and eliminate the non-stationary part of the time series. This differencing also eliminates components as trend and seasonality, which is also important when using the Moving Average model.

$$y'_t = y_t - y_{t-1}$$

Figure 44 - Differencing Time Series

The difference is as simple as the equation of Figure 44. Sometimes, after one differentiation the stationary behaviour is not quite achieve. If this happen it would be required to do another differentiation, which is called second order differencing.

There is also a distinction between Non-seasonal Arima models and Seasonal Arima model, which summarized, means that there are two approaches for this type of model. If the time series has a seasonal behaviour we have to choose our order of each model based on the number of periods each season has. The time series we would use for this approach can be found in Figure 45.

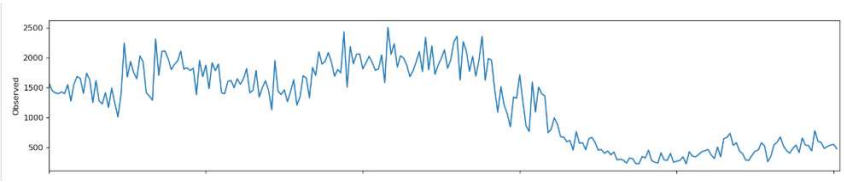


Figure 45 - ARIMA time series

First, we would decompose the time series for us to see the parts that are inside of it. This would give us some information about the procedures that we have to make in order to treat correctly all the characteristics of the time series at hand. From this decomposition we would move forward to studying the degrees in the previously explain models inside the Arima.

A simple decomposition of the time series at hand can be found in Figure 46.

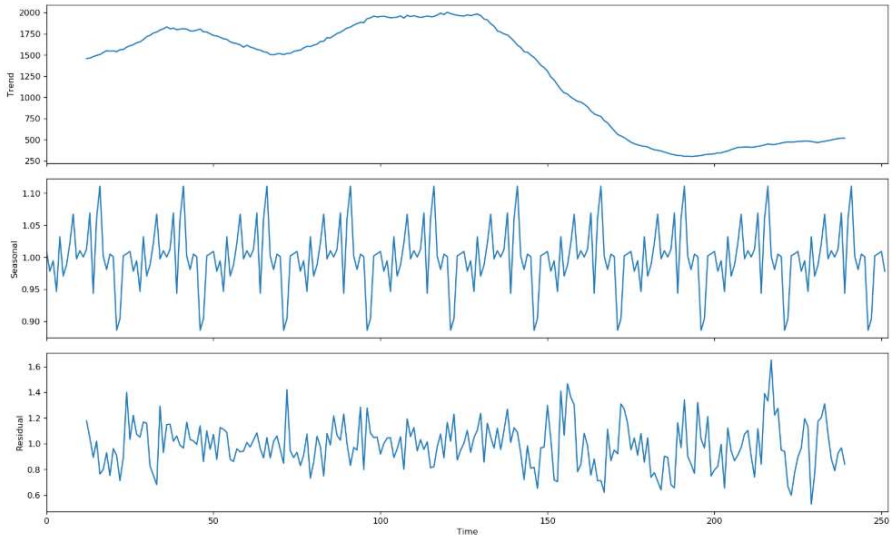


Figure 46 - Time series Decomposition

We can observe how the three main parts of the time series are plotted in the previous image. We have a trend that is mostly stable until the last 100 values where it takes a decreasing turn. After that is begins to increase but in a smaller amount. The seasonality is perfectly extracted by this module of python called: *seasonal_decompose.* This seasonality would have a great influence in our model. We would study the possibility of using a Seasonal Arima for this time series since the seasonality is so present in the actual values. In third place, we have the residuals or noise. This values are not explained either by the Seasonality nor the Trend. This is the randomness of the time series. The Arima model would be focused on treat this residuals until the can follow a Gaussian distribution. That would be the main goal for the next steps in this chapter.

One of the most important parts of the Arima model is to choose the degrees for each one of the models that are working inside of it. We will have three degrees of freedom: p, d and q. The p would be the one associated to the AR model, the d to the Integration/Differentiation part and the q with the Moving Average method. We will have to tune this values in order to achieve a good performance in our model.

The objective of this tuning is to achieve the characteristics that are needed in order to be able to treat the time series. The main aspect of a time series is the existence correlation between the values and time. We have to make disappear this relation so a couple of theorems can be applied to the datasets. Just like the central theorem limit or linear regression. For achieving this stationary behaviour we will focus our efforts in differentiating the time series, we will measure the degree of success based on two plots: ACF and PACF[15].

The ACF plot is correlated with the MA terms. This is explained given that the Moving Average process would be a process with a lagged value until the k

iteration. Since what we do is adding all the values of the precedent values an averaging them creating a vector of averages. Each average would be strongly correlated with the lagged values.

The PACF will be in relation with the AR terms. Defining the correlation for an observation and its lags values.

The initial observation of the time series that we are using to create an Arima model can be seen in Figure 47. Here we can observe the initial values of the time series and an initial correlation plot. This plot would show the correlation of the values the time series with each precedent values, or lags.

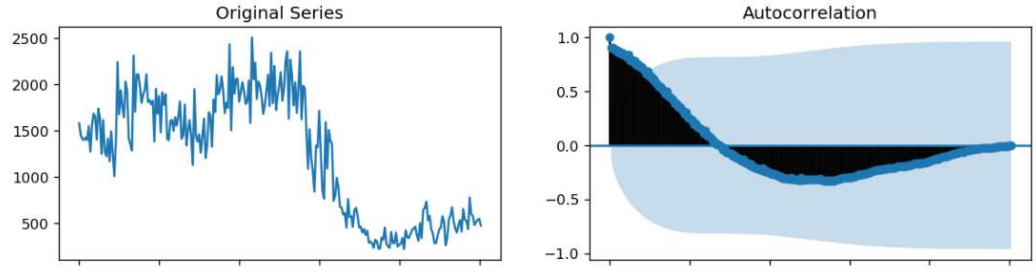


Figure 47 - Original Time Series

The correlation method used for creating this plot by the *statsmodel* library of python is going to be the Pearson correlation index. That's why the values are shown between 1 and -1. The Pearson correlation value give a value that represents the linear relationship between the values. We can see the evolution on the original series and therefore the evolution of the correlation plot, since there is a peak around the year 2009. The correlation plot goes from having a value closed to one to having a negative value. The blue zone that we can observe in the correlation plot is the confidence zone, which is normally set to a 95% interval.

We have to calculate some differencing of this original data set to create a stationary time series. We can observe in the next figure that after one differencing we achieve a better behave but it is not quite perfect.

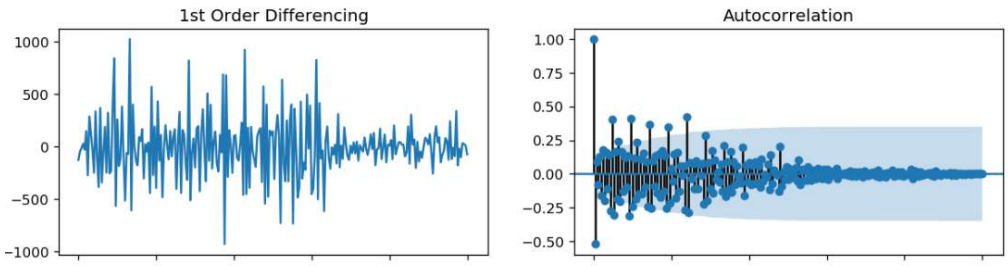


Figure 48 - First Order Differencing

We can also see that there is some seasonality behaviour in the ACF plot. This is not good since if we find some seasonal behave we would have to use the Seasonal Arima, which is not by definition a bad thing but we should always try to simplify our time series as much as we can with the tuning of the parameters with the goal in ind of eliminating all the odd behaviour that make the treatment of the time series more complicated.

In figure 49 we can observe the output after a second differencing. The results are really similar with the previous one, we would stay with this second order differencing for now, keeping in mind that after some testing we can find out that this second differencing is unnecessary and therefore we would have to delete it.

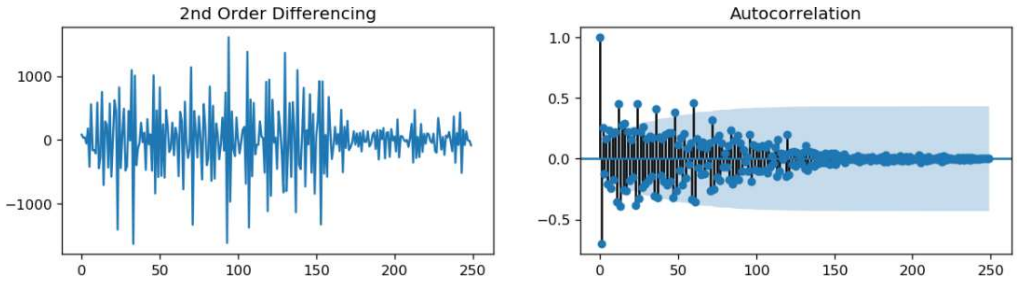


Figure 49- Second Order Differencing

We can also observe the seasonal behaviour in the correlation plot. We haven't achieve a good solution for this seasonal characteristics so we should consider using a seasonal Arima to take this in account.

Now that we have achieve a seasonal state with the ACF plot, it is time to study the PACF plot.

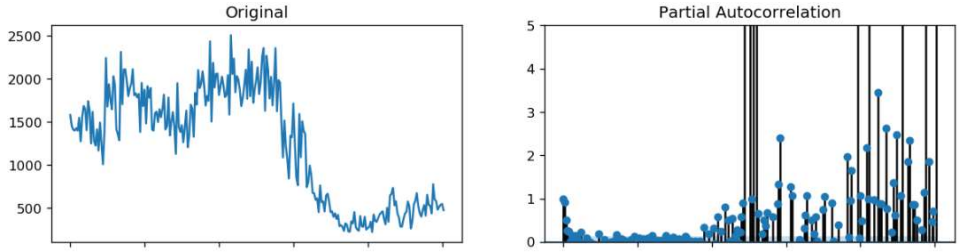


Figure 50 - Arima PACF

This plot gives us a lot of information about the correlation of the values inside the time series. We have to remember that the PACF does a study of the correlation taking in account all the lag values between the two values is correlating. So, for example in Figure 48, we can see that all the values until approximately the value 100 are correlated with the value in k =1. This makes a lot of sense since we have a strong change in the behaviour of the sales around that point. We can obtain different lectures from this plot.

Having a large spike at the first lag that has decreased on the next lags means that we are in the presence of a moving average term in the data that has to be taken care of. We have to find the degree of this moving average using the autocorrelation function.

For the other values that are shown in the lags that are far away from lag number 1, we will differentiate the function to find what the reason of its behaviour is.

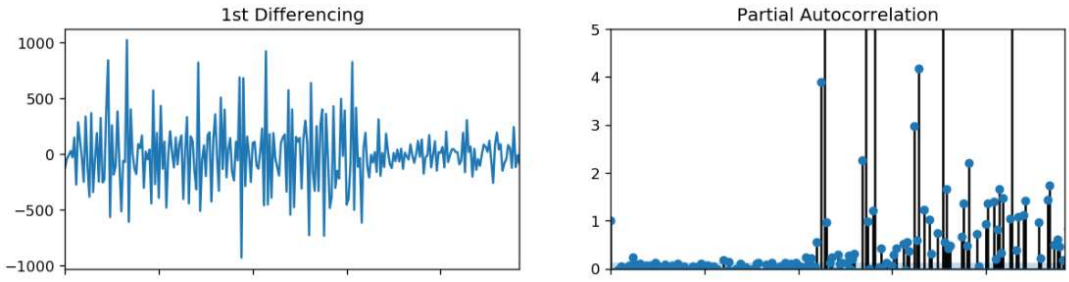


Figure 51 - Arima PACF 1 order

We can observe how the lags have decrease on the volume until the point we have the unusual behaviour appears. We can observe that this correlation has some kind of repetitive behavior which could be explained by a seasonality component in the time series. If a seasonality component appears in our time series that would be there is a higher order moving average component present in the time series and we would have to use the correlation function again to determine its order. If we take a second differentiation to see a third plot and extract more info we can find interesting information.

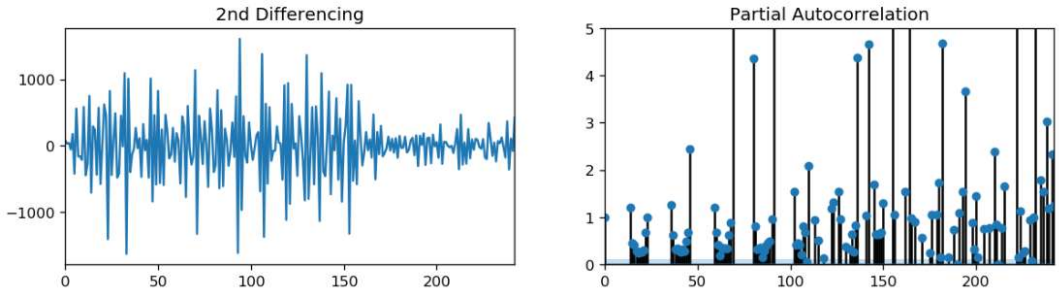


Figure 52 - Arima PACF 2 order

In Figure 52 we can observe a really characteristic seasonal behaviour, we can attack this problem by two means. Either we use a seasonal Arima model that takes care of this behaviour or we try to differentiate the time series enough so the seasonal behaviour disappears. Using the Arima seasonal model looks like a better approach since the pattern of a seasonal characteristic is clearly present.

From the same python package we can also get a summary of the chosen model. Where we can see the importance of the coefficients chosen with the differentiation order. The main value we have to check in this summary is the P_value. We will show first the summary created using an Arima model with these characteristics:

$$\text{model} = \text{ARIMA}(\text{TimeSeries}, \text{order}=(4,2,1))$$

We have set the values (4,2,1) for the order of each model inside Arima based on the analysis done in the previous part, analyzing the PACF and the ACF plots. If the p_value is smaller than 0.05 then that coefficient is important for the model and we should keep it. If the p_value is excessively big, that coefficient can be disregarded from the model. In Figure 51 we can observe an example of this summary for the present model.

ARIMA Model Results

Dep. Variable:	D2.y	No. Observations:	250
Model:	ARIMA(4, 2, 1)	Log Likelihood	-2117.610
Method:	css-mle	S.D. of innovations	1150.672
Date:	Thu, 18 Jul 2019	AIC	4249.219
Time:	14:12:25	BIC	4273.870
Sample:	2	HQIC	4259.140

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1556	20.136	-0.057	0.954	-40.621	38.310
ar.L1.D2.y	-1.5309	0.138	-11.121	0.000	-1.801	-1.261
ar.L2.D2.y	-1.3443	0.164	-8.207	0.000	-1.665	-1.023
ar.L3.D2.y	-0.8278	0.127	-6.526	0.000	-1.076	-0.579
ar.L4.D2.y	-0.3675	0.061	-6.030	0.000	-0.487	-0.248
ma.L1.D2.y	0.3954	0.141	2.801	0.005	0.119	0.672

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.0789	-1.3687j	1.3710	-0.2592
AR.2	-0.0789	+1.3687j	1.3710	0.2592
AR.3	-1.0472	-0.5924j	1.2031	-0.4181
AR.4	-1.0472	+0.5924j	1.2031	0.4181
MA.1	-2.5289	+0.0000j	2.5289	0.5000

Figure 53 - Summary ARIMA

We can observe how the first value of the Moving Average model is a little bit bigger than for the rest of the coefficients. We can also observe how the rest of the coefficients are explaining a great amount of the behaviour of the data a therefore they would be important when we fit the model. All this modeling has one goal: to achieve an stationary behaviour of the residuals. We have to have a density of the residuals that can be modeled by a Gaussian distribution. This condition is key for all the assumptions we have make in order to create the forecast. If the residuals are following a Gaussian we can, for example, apply the Central Theory Theorem. We can observe, in the next Figure, how the model hasn't quite achieved yet a stationary behaviour for the residuals.

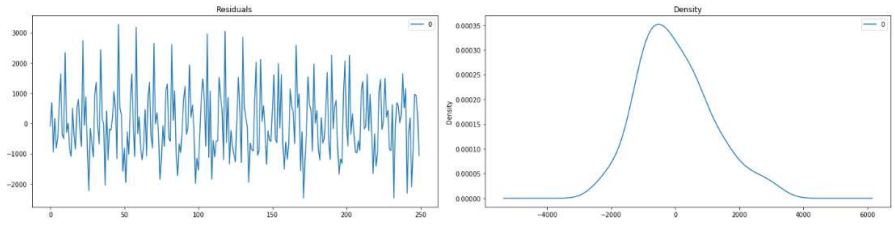


Figure 54 - Residuals Distribution

We can observe how we have achieve a close behaviour to the optimal, but also how the residuals are not totally centered in the 0. This could led to problems with the prediction so we should try to increase the accuracy of the prediction so the residuals are following the Gaussian distribution as they should. With a model with the orders (3,2,1) we obtain the following residuals:

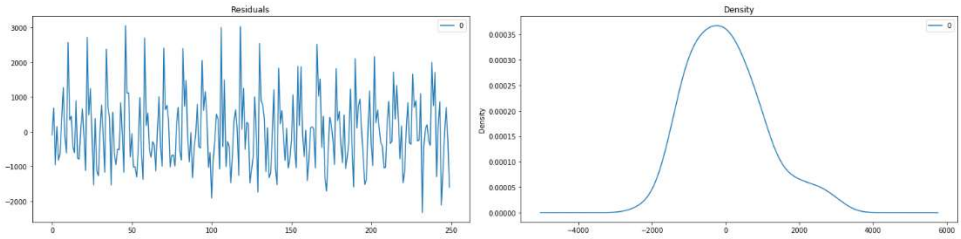


Figure 55 - Residuals Distribution adjusted

The tendency to the left persists but the effect has decrease, therefore, we have achieved a better model eliminating the 4th degree from the Autoregressive model, as the summary in the table in Figure 53 showed. We can also plot the prediction made with the model in order to calculate this residuals. We can observe that in Figure 54.

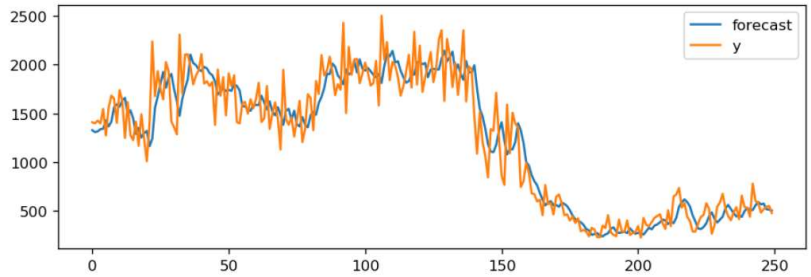


Figure 56 - Forecast prediction

As we can see the model is good but not perfect for an accurate forecast. We can work with this model to create a preliminary forecast and see how this model would adapt to this data set. In order to validate it, we would cut the last values of the time series and predict them. Then, we will plot both, the real and the forecast, last values in order to see the difference between them. If the next Figure we can observe this first prediction on the last 50 values of the time series.

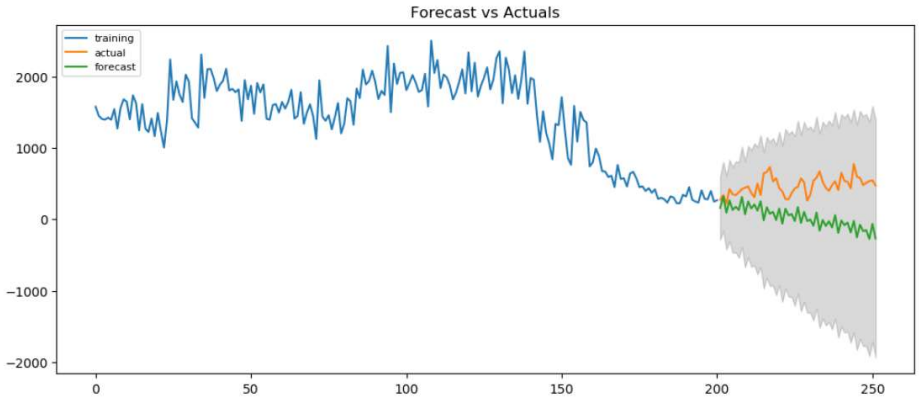


Figure 57 - Forecast implementation

As we can see. The prediction is far to be good enough. We can observe how the predicted trend is going way below the actual values. This could be explained because we haven't achieved a perfect seasonality behaviour and therefore the residuals are influencing too much in the future values. A way to increase the accuracy with our model, would be to add the residuals of that values to see how the prediction would be taking them in account. If we have a closer prediction, we would have to create a new model that handles better the residuals in order to get a more accurate prediction.

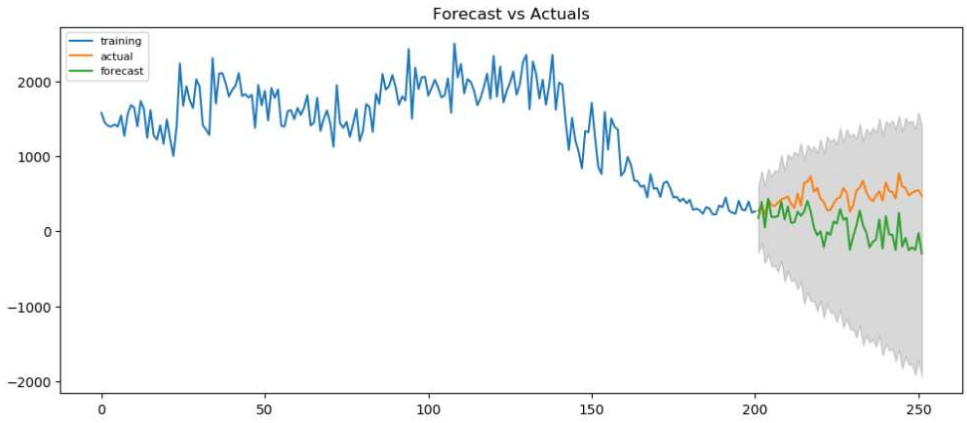


Figure 58 - Adjusted prediction

Even if this is not the best approach to increase the accuracy of our model it is a great way of showing that we have to take more in account the residuals present in the model since they are not fully taken in account in the prediction. The figure presents a more decent behaviour for our forecast. Even if we don't achieved a perfect prediction, after adding the residuals we obtain a very similar behaviour, which is a good starting point. The grey area we can see in the plot makes reference to the area of confidence of the forecast. If our forecast is inside the grey area we can say that we have a 95% confidence on that prediction.

Even though we were achieving interesting results in the Forecast implementation, they business department decided that we had to continue in other directions. We had to go back to the correlation problem. The Forecast wasn't exactly what they were looking for, since they didn't have interest in the value of the future, but instead, on the global behaviour of the total demand time series.

At this point, we decided which type of solution we were going to make. The idea would be to create a tool that would give three key outputs:

- Matrix Correlation
- Direct – Inverse plots
- Summary table

We will see the actual implementation of these three calculations in the *Results & Conclusion* chapter. In order to achieve this solution, first we had to attack another problem. The categorization of the total demand for each country.

k. Categorization

The categorization was an issue that was assessed once the tool and its solution was already decided. Once we already knew that we had everything necessary to start creating some wisdom we decided to close the actual scope of the project in order to achieve for meaningful insights. We decided to start a categorization between the Light duty trucks and the Heavy-Medium duty trucks. Once the categorization of the variables would be finish we would decide the countries were we would focus the study. This part is strongly correlated with the *Cleaning & Structuring* chapter of the Introduction.

In the next figure we can observe an example of the type of document we obtained once we started with the categorization problem. This Figure is for the Light Duty vehicles.

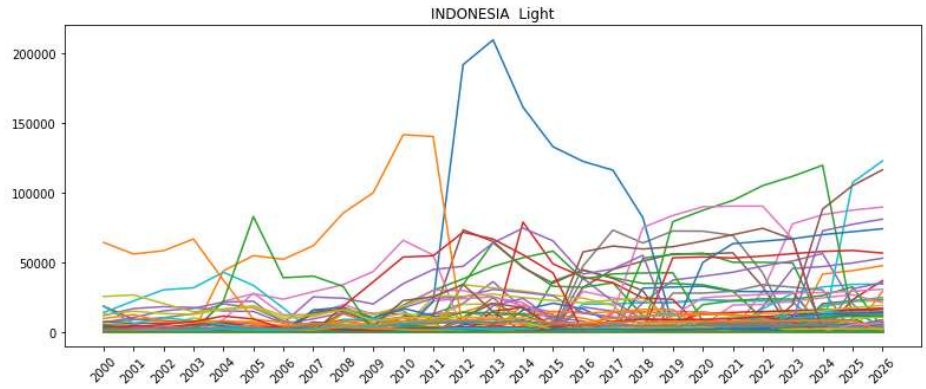


Figure 59 - Indonesia Light

As we can see, we have data until the year 2026, this is because we will be using this forecast in order to categorize the countries in a more reliable way. Also we would have to remark the random structure of the demand of the light commercial vehicles. This is explained by the different types of vehicles that are included in this category. We have from cars to light duty trucks. This data has to be filtered to obtain only the light duty trucks that are the object of study of the present project. We have also to study the possibility of filtering also the vans, since there are some countries that do not include the vans inside of this category. However, the changes in the total demand of vans, can affect the total demand of the light duty trucks. Therefore, we can't filter completely the effect of this category on the variable that we are studying.

We have followed the categorization done by another team inside Daimler, in order to be able to compare the conclusion we gain in this project with the conclusion obtained by the *best finance team*. Because of that, we have filtered the cars for this *light* document and leave the vans, as for now. In the next Figure we can observe the distribution of the other file taken in account in the categorization problem.

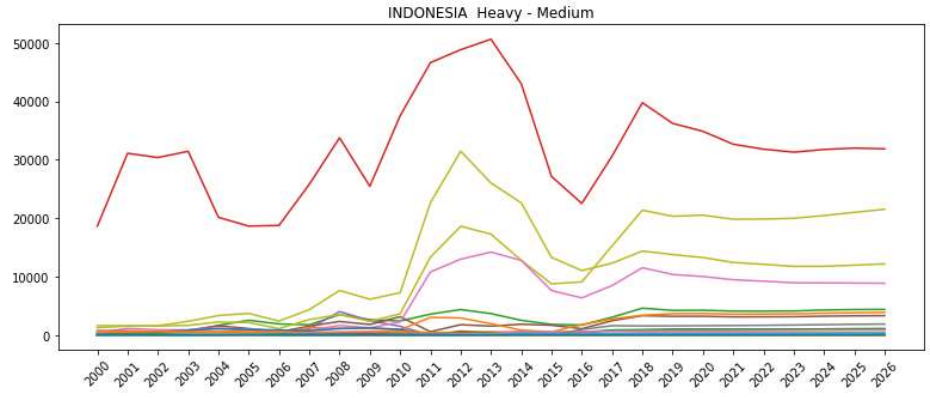


Figure 60 - Indonesia Heavy-Medium

The comparison between the previous images is clear. We can observe the difference between the two plots. In the Heavy-Medium categorization we have a common distribution for almost all the branches. This is normal, since the categorization for the Heavy duty trucks is simpler. All the vehicles recorded here have more than 15 ton. Which makes that the correlation between branches is strong. The market will influence in the total demand of all the branches similarly. This is way it would be really easy working with this categorization rather than with the Light duty vehicles. The conclusions obtained by the tool for this categorization would be global for all the branches and we could apply its results for all the heavy-medium companies in Indonesia.

Since we have to use this categorization for the tool that would be created by the department, we have to confirm that the data that we have is reliable in comparison with the categorization take by the *best finance team*. We would filter the values from this two document with the same rules that the *best finance team* used for categorize its data. After that we would compare the result with the data we were working and compare if the structure, but foremost, the evolution over time, is similar enough. If we have similar values, we would consider the possibility of continuing with our hypothesis. If we obtained a different structure for the total demand. We can use the data obtained by the combination of this two source for the study.

This comparison will also help us to see the lack of homogeneity in the sources. Since, as we would find out in the next figure, the numbers given by different sources may be similar in the evolution but they have a huge different

regarding the volume of the total demand. Since the branches not always display the actual values of its sells, we will get a huge gap between the actual values that are recorded by OICO and the aggregate of all the values from each brand.

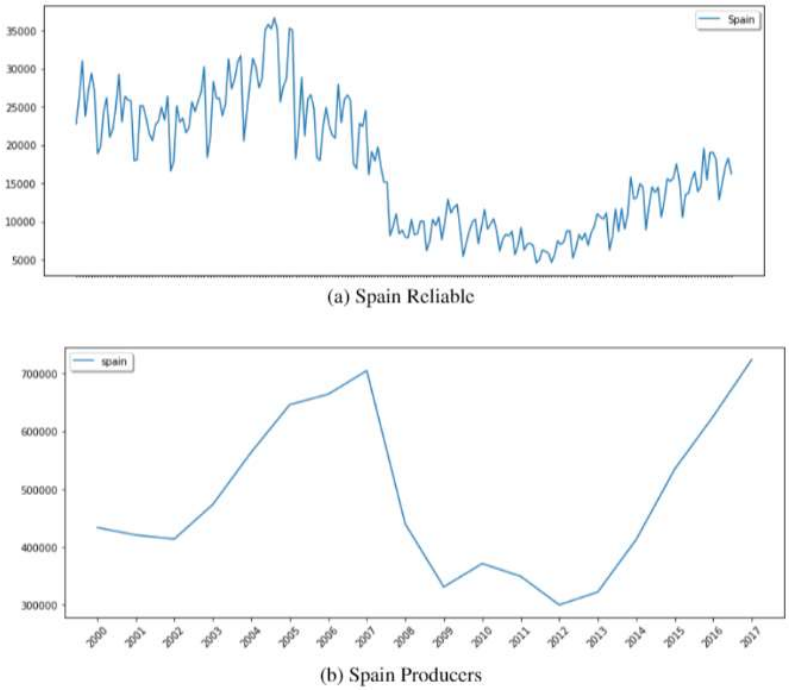


Figure 61 - Comparison of sources

As we can see, the numbers given by the producers don't add for the official values given by the reliable sources. It is because of this that this study is interesting, since we have to understand how the total demand is going to behave in the future without rely in the data given by the producers. Apart from this, we can see that the evolution is similar, for the official source, we only have values until 2017, but we can see that the evolution over time until that point is similar. We have some differencing since the values of the producers are recorded monthly and we can see more fluctuation, but the evolution, which is what matters at this point, is comparable, and we can continue with our study without relying on the values given by the branches.

Before showing which approaches were finally taken and show the results obtained by this project, we will present another approach that we study as a possibility for future endeavors. Since all the variables that we have so far are economical, we wanted to take also in account the variables that can be measure with numbers but that they would also influence in the total demand. This approach was called Event Correlation.

l. Event Correlation

Other approach regarding this project was concerning the impact of the events on the total demand, such as political elections or the approving of environmental laws concerning the prices of the fuel. This approach was focused for countries with a low development structure since they are the most influenced by this changes. For example, in Africa, it is known that during the electoral campaigns most of the demand is stopped because of the incertitude before the elections. The idea was to try to find the breakpoints in the time series and then trying to find the cause of this breakpoint using the data of the different event that we had gathered. This approach couldn't be finished since we focused more on the correlation approach, but we did achieve some interesting results finding the breakpoints of the time series using package ruptures from Python.

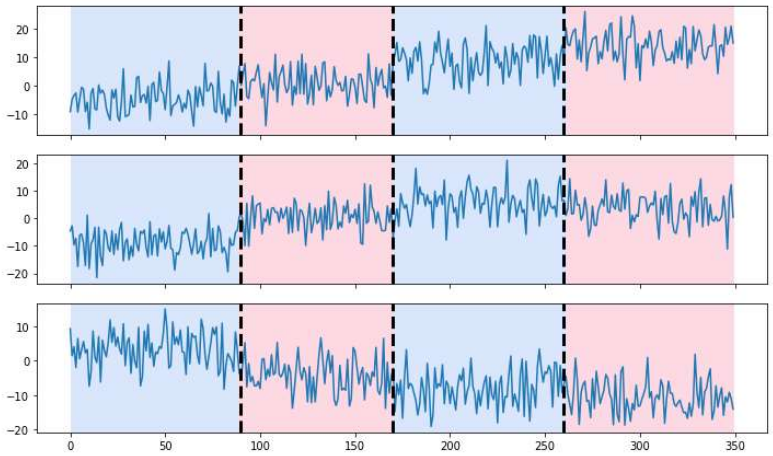


Figure 62 - Event correlation example

In figure 62 we can observe an example of the usage of this package. This package studies the evolution of the time series and tries to identify the different breakpoints that can be found on it. From that breaks the time series in different smaller time series. It is possible to set the number of breakpoints we are looking for and also the algorithm used for the calculations. The idea behind this procedure is that at some point the trend and the seasonality of the time series changes so hard that it is possible to identify as a different time series. Even if, for the human eye, the behavior of the time series is similar through all the time spectrum, this package can identify the differences between the different data and break the time series. This would be the first part of this approach. The second one should compare these breakpoints found by this algorithm with the events that we have gathered for ourselves. This approach has a lot of flaws since the breakpoints are founded by the algorithm and, for now, we can't set what are the changes that we are looking for. Also, we may be looking at different events that are not correlated with the breakpoints found by the algorithm. This is why we did the procedure the other way around. Instead of trying to found the breakpoints that we are looking for in the time series, we let the algorithm tells us where are the breakpoints and then we compare those with our events, therefore we reduce the possibility of finding some results that are not quite correct but since we are looking to find some results we would be eager to find breakpoints where they are not. Using this procedure, we reduce the possibility of error but also the possibility of getting results easily. For this method to be useful we should gather a big number of events in order to be able to compare them with the breakpoints. To this we can add the complexity of the countries. At some point, one event can be significant to one country but not to another, therefore we cannot extrapolate the results of one country to another. For this, we should try to find some kind of clustering that can be useful in order to find the events that are influencing for the same countries. This clustering can be geographical of financial. At this point, the project followed another methodology. We did get some interesting results obtaining the breakpoints of one of the countries involved in some of the events we have gathered. We can observe its evolution in the next Figure.

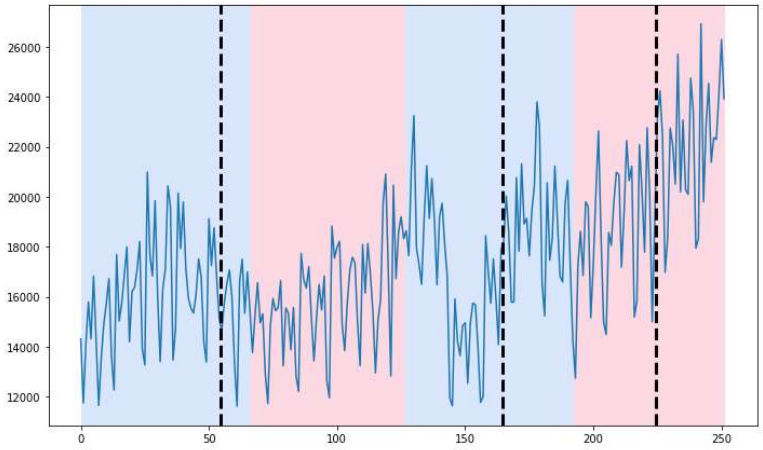


Figure 63 - Event correlation implementation

As we can observe, we are able to find four breakpoints for this time series. This time series belong to the total demand of commercial vehicles of Germany. We can see how the colors do not match the lines, this is because we have to make some arrangements with the algorithm since it is not quite robust if we want to change the amount of values the time series has or the number of breakpoints we are looking for. This is another problem regarding this algorithm, we should think about another procedure in order to decided how many breakpoints we are looking for. There is another package in Python where it finds the hypothetical breakpoints without the predefined number by an user. If this procedure works perfectly, we should focus on that one since defining how many breakpoints we want to find can biased the decision of the algorithm and therefore we can be founding some results but because we are inducing the algorithm to find them for us, even if there is not breakpoint there.

As we can see, this method has a lot of potential but needs a lot of time to be able to be useful somehow, it is because of that that we didn't continue working on this method during the last period of the project since we had to focus of the correlation analysis since that one was more likely to be finished on time. For future developments for this methodology we should gathered a big database of events we want to study, we should create some methodology to cluster the countries with a similar behavior concerning this kind of events and we should improve the rupture package in order to be able to find the breakpoints that are present in the time series without the input of an user. If those three things can be achieved, we should be able to find some interesting results with this procedure and therefore understand better how the behavior of the countries regarding the events is. Once we have achieve this we can plan our strategy based on this events, and therefore, if for example, a big country

has a new environmental law changing the prices of the fuel and we saw that on previous similar occasions this has influenced negatively on the evolution of the total demand of commercial vehicles for the country, we can assume that that behavior is going to be repeated and therefore we can start some movements in order to minimize the impact in our sales.

Once we have already listed all the methods studied during the time of this project. We should move forward to the actual results and final methodology that we created during this period. This methodology is based on the correlation study seen in previous chapters. We would do a summary of the main aspects of this procedure and then explain what was the actual delivery that we handed to the Financial colleagues.

6. Results & Conclusion

In this chapter we will describe the final deliverable we gave to the Financial Team. Since we see all the different approaches we tried to study, we will summary what was the final result. The final result was based on the correlation analysis we explained in previous chapters. This was because of the needs of the department. We didn't move forward in other approaches like DTW and DBA since there complexity was large enough and we couldn't finished its implementation with the time we had. Therefore, in order to meet the department needs we all agreed that the best solution was to give an easy and self explanatory deliverable. Therefore, the final results is a summary of the correlation analysis, taking the three main results from its study: the correlation matrix, the correlation table and the correlation plots. We created results for all the countries listed in that chapter for all the variables selected for the study.

Countries:

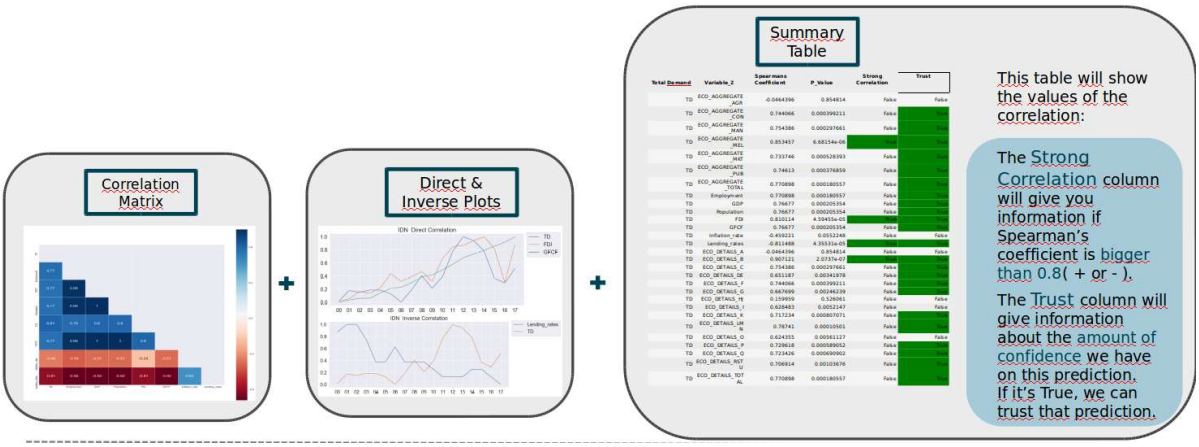
- Indonesia
- Hong Kong
- China
- Vietnam
- Thailand
- Malaysia
- Philippines
- Singapur
- Taiwan

Variables:

- Gross Domestic Product

- Gross Value Added
- Direct Foreign Investment
- Interest Rates
- Employment
- Population
- GFCF

Therefor, the deliverable was a document with all the insights for all these countries. We increased the scope of the project using all the sectors inside the Employment variable. The next step was to make the study for countries based in Africa where the confidence for the data of the total demand was smaller. In the next figure we can see how the end result would look like.



Therefor, the final result would give the Financial team fast insights, reliable information and gain of knowledge. This, inside the big scope of the financial team, should be really helpful for localizing the most important variables regarding the total demand of commercial vehicles. Given that the main goal of this project was to categorize the countries into rising markets or declining markets this results gives a good starting point to continue to categorize the countries with a data driven point of view.

We can summarized the main aspects of this projects, or conclusions, in three main points:

- Knowledge was gained
- A solid base was developed for future endeavors
- A deliverable was finally handover to the Financial colleagues

For future developments, we should try to focus our efforts in finding more reliable sources where to get more useful information. For that is already established a solid base where, when the enough amount of data is stored, really interesting projects can take place.

As a personal conclusion, the project was interesting and challenging. Making myself to make a great effort to be able to handle the importance of the project.

7. Resources

For this project we have used mostly Python. In the next section we will enumerate all the packaged used during this project. For using Python we used Jupyter notebook for the development environment. We have also used Github for storing all the files. The link for the github repository is:

https://github.com/Salinatorimperator/NDA

a. Python

We used Python as the main language for this project, all the calculation were made using Python packages and its building-in functions:

i. Pandas
    1. Std
    2. Loc
    3. Apply
ii. Sklearn
    1. PCA
    2. Standard Scaler
    3. kmeans
iii. Numpy
iv. Matplotlib
v. tslearn

We also used some classes like the SSA class object for an easier calculation method.

b. Jupyter

We used Jupyter for all the calculations and processing of the files. We created some .py files to automatized the process but since none in the Financial Deparment knew how to run a .py. We stayed with Jupyer. In the repository we can see one file called “Pascal Data Visualization”, this file was created so Pascal, from the Financial Team could experiment with the code and do some changes.

8. Annexes
    a. Code

Code for obtaining the Final Result for a country:

```
#Choose country
country = 'PHL'

# Correlation Matrix

line = copy.loc[copy['Countries'] == country]
val=pd.DataFrame()
for column in line:
   val[column] = line[column].values[0]

val = val.drop(columns=['Countries'])
val['TD'] = pd.to_numeric(val['TD'])



mask = np.zeros_like(val.corr())


mask[np.triu_indices_from(mask)] = True
colormap = plt.cm.RdBu

plt.figure(figsize=(30,30))

#We create the correlation matrix with all the data
sns.set(font_scale=1.8)
svm = sns.heatmap(val.corr(method = 'spearman'), mask=mask,
linewidths=0.1,vmax=1.0,
```

```
        square=True, cmap=colormap, linecolor='white', annot=True,
annot_kws={"size": 30})


# Plots

val_normalized = val[:]
min_max_scaler = preprocessing.MinMaxScaler()
val_scaled = min_max_scaler.fit_transform(val.values)
val_normalized.loc[:,:] = val_scaled



col_names = val_normalized.corr().columns.values
years = ['00','01','02','03','04','05','06','07','08','09','10','11','12','13','14','15','16','17']
figure(figsize=(13,5))
plt.title(country + '  Direct Correlation')

for col, row in (val_normalized.corr(method='spearman') > 0.80).iteritems():
   if col =='TD':
      for c in col_names[row.values]:
            plt.plot(years,val_normalized[c], label = c)
#plt.legend(shadow=True, fancybox=True)
plt.legend(bbox_to_anchor=(1.0, 1.05))


figure(figsize=(13,5))
plt.title(country + '  Inverse Correlation')

plt.plot(years, val_normalized['TD'], label = 'TD')
for col, row in (val_normalized.corr(method='spearman') < -0.50).iteritems():
   if col =='TD':
      for c in col_names[row.values]:
            plt.plot(years,val_normalized[c], label = c)

#plt.legend(shadow=True, fancybox=True)
plt.legend(bbox_to_anchor=(1.0, 1.05))



# Table with P_value
col=[]
```

```
results = pd.DataFrame()
j=1

for column in val:
   col.append(column)

variable1 = val[col[0]]
name1=col[0]
for i in range(0,len(col)):
      if i == len(col)-1:
         break
      else:
         variable2 = val[col[i+1]]
         name2=col[i+1]
      corr, p = spearmanr(variable1, variable2)
      results.loc[j, 'Total Demand']=name1
      results.loc[j, 'Variable_2']=name2
      results.loc[j, 'Spearmans Coefficient']=corr
      results.loc[j, 'P_Value'] = p
      j+=1
results['Strong Correlation'] = results['Spearmans Coefficient'].apply(lambda x: abs(x)>
0.8)
results['Trust'] = results['P_Value'].apply(lambda x: abs(x)< 0.005)


results.style.apply(highlight_results, subset=['Strong Correlation', 'Trust','Spearmans
Coefficient'])
#results.style.apply(highlight_negative, subset=['Spearmans Coefficient'])
```

Output:

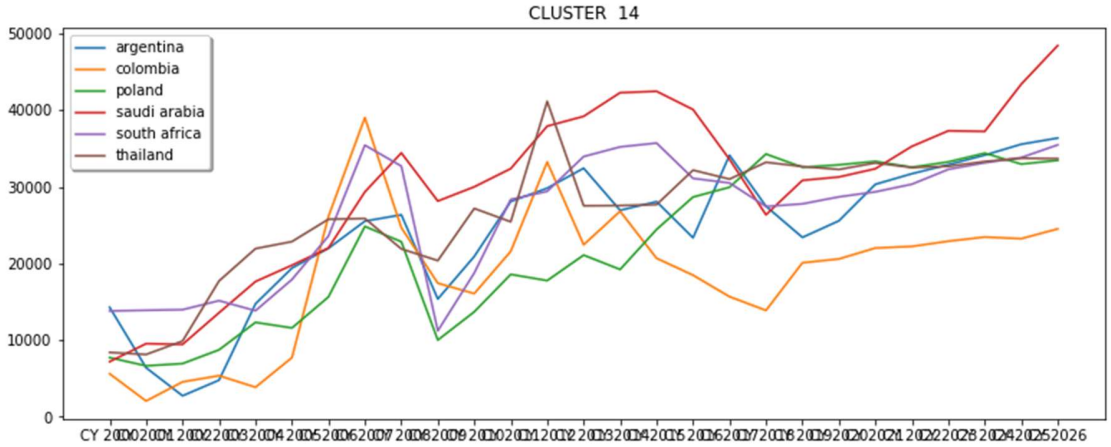
The output of this code is in the results Chapter.

Code for clustering countries:

```
kmeans=KMeans(n_clusters=15,random_state=0).fit_predict(total.drop(columns
='Countries').values)
kmeans

cluster = 14
k_1 = kmeans==int(cluster)
plot_data = total.loc[k_1]
```

```
Result = plot_data.loc[:, plot_data.columns != 'Countries'].values

Date = list(plot_data)[1:]
#Result_4= list(Result.reshape(13,len(data.loc[k_1])))
labels = list(plot_data['Countries'])
figure(figsize=(13,5))
plt.title('CLUSTER  '+ str(cluster))

for i in range(0,len(Result)):
   plt.plot(Date, Result[i], label=labels[i])

plt.legend(shadow=True, fancybox=True)
```

Output:



Class for Singular Spectrum Analysis:

```
class SSA(object):

   __supported_types = (pd.Series, np.ndarray, list)

   def __init__(self, tseries, L, save_mem=True):
      """
```

```
    Decomposes the given time series with a singular-spectrum analysis. Assumes the values of the time series are
    recorded at equal intervals.

    Parameters
    ----------
    tseries : The original time series, in the form of a Pandas Series, NumPy array or list.
    L : The window length. Must be an integer 2 <= L <= N/2, where N is the length of the time series.
    save_mem : Conserve memory by not retaining the elementary matrices. Recommended for long time series with
        thousands of values. Defaults to True.

    Note: Even if an NumPy array or list is used for the initial time series, all time series returned will be
    in the form of a Pandas Series or DataFrame object.
    """

    # Tedious type-checking for the initial time series
    if not isinstance(tseries, self.__supported_types):
        raise TypeError("Unsupported time series object. Try Pandas Series, NumPy array or list.")

    # Checks to save us from ourselves
    self.N = len(tseries)
    if not 2 <= L <= self.N/2:
        raise ValueError("The window length must be in the interval [2, N/2].")

    self.L = L
    self.orig_TS = pd.Series(tseries)
    self.K = self.N - self.L + 1

    # Embed the time series in a trajectory matrix
    self.X = np.array([self.orig_TS.values[i:L+i] for i in range(0, self.K)], dtype ='float').T

    # Decompose the trajectory matrix
    self.U, self.Sigma, VT = np.linalg.svd(self.X)
    self.d = np.linalg.matrix_rank(self.X)

    self.TS_comps = np.zeros((self.N, self.d))

    if not save_mem:
        # Construct and save all the elementary matrices
```

```
        self.X_elem = np.array([ self.Sigma[i]*np.outer(self.U[:,i], VT[i,:]) for i in
range(self.d) ])

        # Diagonally average the elementary matrices, store them as columns in array.
        for i in range(self.d):
            X_rev = self.X_elem[i, ::-1]
            self.TS_comps[:,i] = [X_rev.diagonal(j).mean() for j in range(-X_rev.shape[0]+1,
X_rev.shape[1])]

        self.V = VT.T
      else:
        # Reconstruct the elementary matrices without storing them
        for i in range(self.d):
            X_elem = self.Sigma[i]*np.outer(self.U[:,i], VT[i,:])
            X_rev = X_elem[::-1]
            self.TS_comps[:,i] = [X_rev.diagonal(j).mean() for j in range(-X_rev.shape[0]+1,
X_rev.shape[1])]

        self.X_elem = "Re-run with save_mem=False to retain the elementary matrices."

        # The V array may also be very large under these circumstances, so we won't
keep it.
        self.V = "Re-run with save_mem=False to retain the V matrix."

      # Calculate the w-correlation matrix.
      self.calc_wcorr()

    def components_to_df(self, n=0):
      """
      Returns all the time series components in a single Pandas DataFrame object.
      """
      if n > 0:
        n = min(n, self.d)
      else:
        n = self.d

      # Create list of columns - call them F0, F1, F2, ...
      cols = ["F{}".format(i) for i in range(n)]
      return pd.DataFrame(self.TS_comps[:, :n], columns=cols,
index=self.orig_TS.index)


    def reconstruct(self, indices):
      """
```

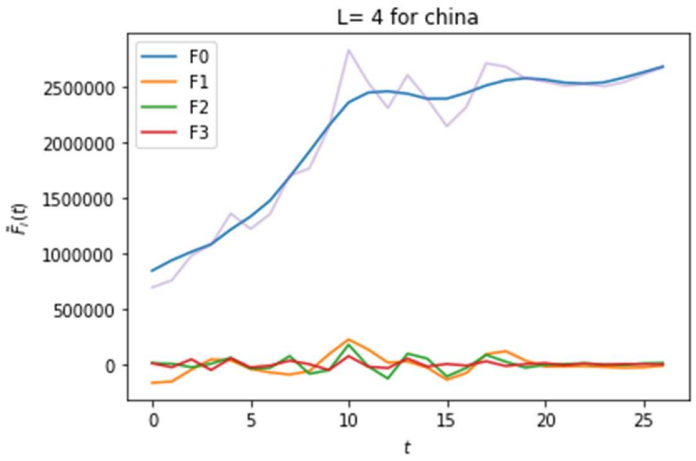
```
        Reconstructs the time series from its elementary components, using the given indices. Returns a Pandas Series
        object with the reconstructed time series.

        Parameters
        ----------
        indices: An integer, list of integers or slice(n,m) object, representing the elementary components to sum.
        """
        if isinstance(indices, int): indices = [indices]

        ts_vals = self.TS_comps[:,indices].sum(axis=1)
        return pd.Series(ts_vals, index=self.orig_TS.index)

    def calc_wcorr(self):
        """
        Calculates the w-correlation matrix for the time series.
        """

        # Calculate the weights
        w = np.array(list(np.arange(self.L)+1) + [self.L]*(self.K-self.L-1) + list(np.arange(self.L)+1)[::-1])

        def w_inner(F_i, F_j):
            return w.dot(F_i*F_j)

        # Calculated weighted norms, ||F_i||_w, then invert.
        F_wnorms = np.array([w_inner(self.TS_comps[:,i], self.TS_comps[:,i]) for i in range(self.d)])
        F_wnorms = F_wnorms**-0.5

        # Calculate Wcorr.
        self.Wcorr = np.identity(self.d)
        for i in range(self.d):
            for j in range(i+1,self.d):
                self.Wcorr[i,j] = abs(w_inner(self.TS_comps[:,i], self.TS_comps[:,j]) * F_wnorms[i] * F_wnorms[j])
                self.Wcorr[j,i] = self.Wcorr[i,j]

    def plot_wcorr(self, min=None, max=None):
        """
        Plots the w-correlation matrix for the decomposed time series.
        """
        if min is None:
```

```
        min = 0
    if max is None:
        max = self.d

    if self.Wcorr is None:
        self.calc_wcorr()

    ax = plt.imshow(self.Wcorr)
    plt.xlabel(r"$\tilde{F}_i$")
    plt.ylabel(r"$\tilde{F}_j$")
    plt.colorbar(ax.colorbar, fraction=0.045)
    ax.colorbar.set_label("$W_{i,j}$")
    plt.clim(0,1)

    # For plotting purposes:
    if max == self.d:
        max_rnge = self.d-1
    else:
        max_rnge = max

    plt.xlim(min-0.5, max_rnge+0.5)
    plt.ylim(max_rnge+0.5, min-0.5)
```

Using this class, and with this code, we can get this results:

```
country = 'china'
F = pd.Series(finalfinal['TD'].iloc[0])


L = 4
F_ssa_L = SSA(F, L)
F_ssa_L.components_to_df().plot()
F_ssa_L.orig_TS.plot(alpha=0.4)
plt.xlabel("$t$")
plt.ylabel(r"$\tilde{F}_i(t)$")
plt.title('L= '+ str(L) +' for ' + country);
```

Output:

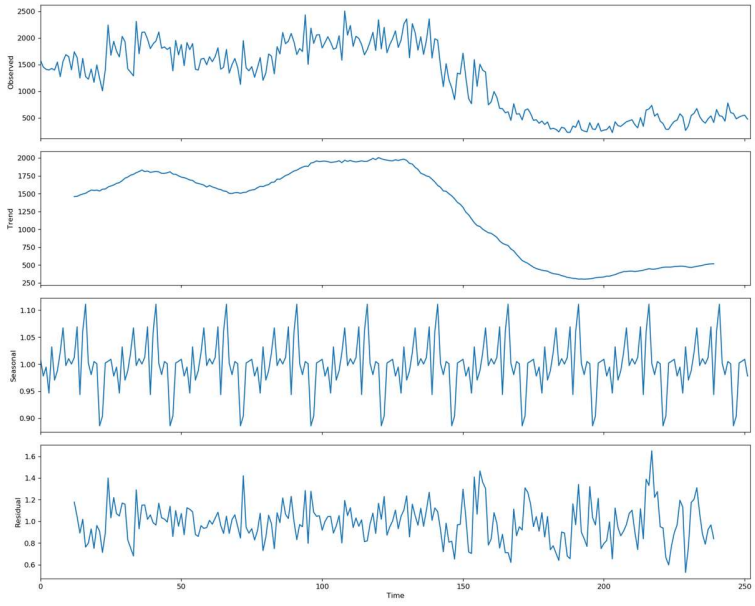


TimeSeries Decomposition

result = seasonal_decompose(TimeSerie.astype(float), model='multiplicative', freq=25)

fig = plt.figure()
fig = result.plot()
fig.set_size_inches(15, 12)

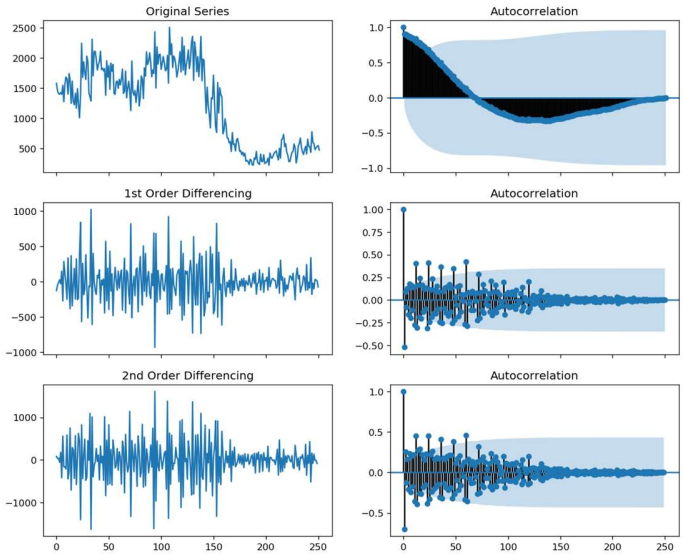
Output:



ACF analysis for TimeSeries:

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
import matplotlib.pyplot as plt
plt.rcParams.update({'figure.figsize':(12,10), 'figure.dpi':120})

TimeSerie = np.delete(df.values[6],0)
# Original Series
fig, axes = plt.subplots(3, 2, sharex=True)
axes[0, 0].plot(TimeSerie); axes[0, 0].set_title('Original Series')
plot_acf(TimeSerie.astype(float), ax=axes[0, 1])

# 1st Differencing
axes[1, 0].plot(np.diff(TimeSerie)); axes[1, 0].set_title('1st Order Differencing')
plot_acf(pd.DataFrame(np.diff(TimeSerie)).dropna().astype(float), ax=axes[1, 1])

# 2nd Differencing
axes[2, 0].plot(np.diff(np.diff(TimeSerie))); axes[2, 0].set_title('2nd Order Differencing')
plot_acf(pd.DataFrame(np.diff(np.diff(TimeSerie))).dropna().astype(float), ax=axes[2, 1])

plt.show()
```

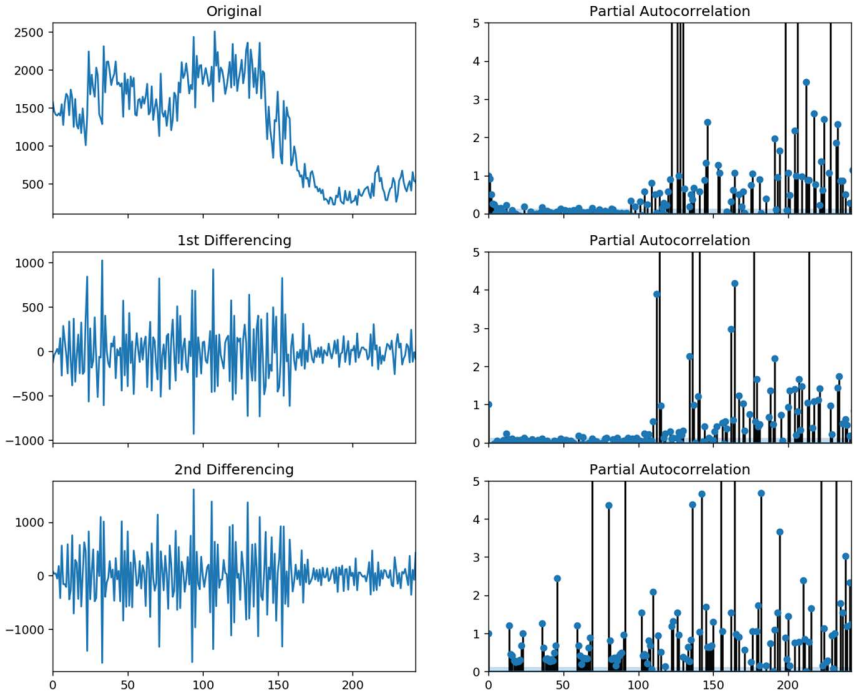
Output:



ACF Analysis:

```
plt.rcParams.update({'figure.figsize':(12,10), 'figure.dpi':120})


TimeSerie = np.delete(df.values[6],0)

fig, axes = plt.subplots(3, 2, sharex=True)
axes[0,0].plot(TimeSerie); axes[0,0].set_title('Original')
axes[0,1].set(ylim=(0,5),xlim=(0,242))
plot_pacf(pd.DataFrame(TimeSerie).dropna().astype(float), ax=axes[0,1])


axes[1,0].plot(np.diff(TimeSerie)); axes[1,0].set_title('1st Differencing')
axes[1,1].set(ylim=(0,5),xlim=(0,242))
plot_pacf(pd.DataFrame(np.diff(TimeSerie)).dropna().astype(float), ax=axes[1,1])


axes[2,0].plot(np.diff(np.diff(TimeSerie))); axes[2,0].set_title('2nd Differencing')
axes[2,1].set(ylim=(0,5),xlim=(0,242))
plot_pacf(pd.DataFrame(np.diff(np.diff(TimeSerie))).dropna().astype(float),
ax=axes[2,1])

plt.show()
```

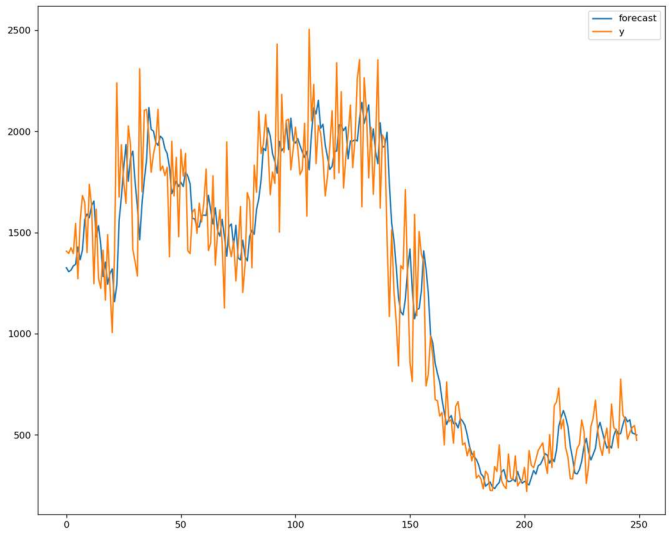
Output:



ARIMA modeling

```
from statsmodels.tsa.arima_model import ARIMA

# 1,1,2 ARIMA Model
model = ARIMA(TimeSerie, order=(3,2,1)) #5,1,3
model_fit = model.fit(disp=0)

# Actual vs Fitted
model_fit.plot_predict(dynamic=False)
plt.show()
```

Output:

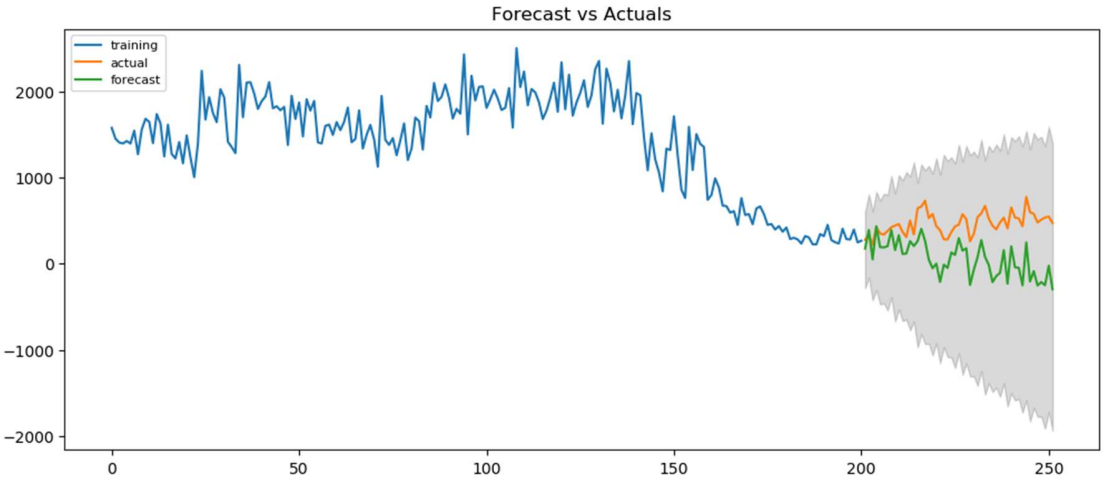


Forecasting:

```
from statsmodels.tsa.stattools import acf

# Create Training and Test
train = TimeSerie[:int(len(TimeSerie) - len(TimeSerie)/5)]
test = TimeSerie[int(len(TimeSerie) - len(TimeSerie)/5):]
# Build Model
# model = ARIMA(train, order=(3,2,1))
```

```
model = ARIMA(train, order=(9, 1, 7))
fitted = model.fit(disp=-1)

# Forecast
fc, se, conf = fitted.forecast(int(len(TimeSerie)/5 + 1), alpha=0.05)  # 95% conf


# Make as pandas series
fc_series = pd.Series(fc, index=pd.Series(test).index + int(len(TimeSerie) - len(TimeSerie)/5))
lower_series = pd.Series(conf[:, 0], index=pd.Series(test).index + int(len(TimeSerie) - len(TimeSerie)/5))
upper_series = pd.Series(conf[:, 1], index=pd.Series(test).index + int(len(TimeSerie) - len(TimeSerie)/5))


test = pd.Series(test, index=pd.Series(test).index + int(len(TimeSerie) - len(TimeSerie)/5))
# Plot
plt.figure(figsize=(12,5), dpi=100)
plt.plot(train, label='training')
plt.plot(test, label='actual')
plt.plot(fc_series, label='forecast')
plt.fill_between(lower_series.index, lower_series, upper_series,
                 color='k', alpha=.15)
plt.title('Forecast vs Actuals')
plt.legend(loc='upper left', fontsize=8)
plt.show()
```

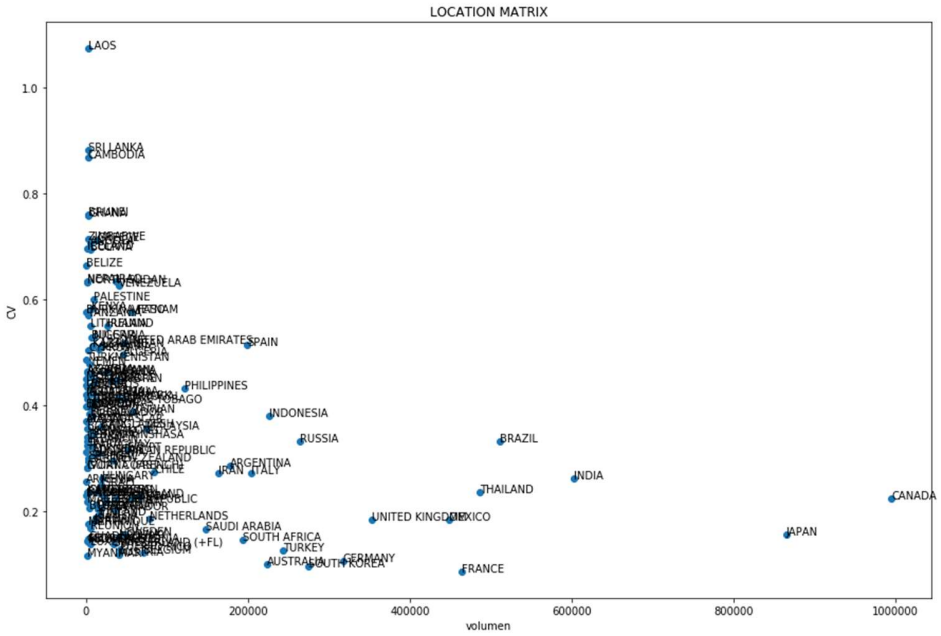
Ouput:



Location Matrix:

```
from matplotlib.pyplot import figure

figure(figsize=(15,10))
ax = display_data.set_index('volumen')['CV'].plot(style='o')
ax.set_ylabel('CV')
ax.set_title('LOCATION MATRIX')
def label_point(x, y, val, ax):
   a = pd.concat({'x': x, 'y': y, 'val': val}, axis=1)
   for i, point in a.iterrows():
      ax.text(point['x'], point['y'], str(point['val']))

label_point(display_data.volumen, display_data.CV, display_data.REGION, ax)
```

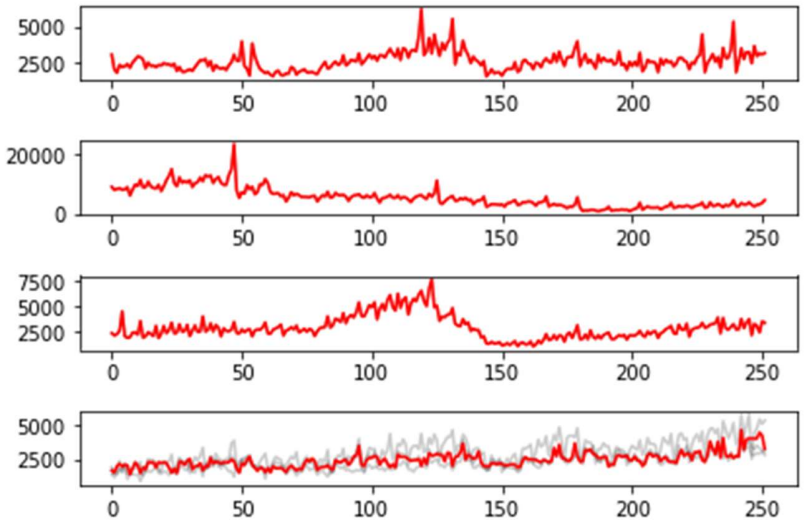
Output:



DTW with DBA:

```
# Soft-DTW-k-means
print("Soft-DTW k-means")

seed=0
cluster =5
k_1 = kmeans==int(cluster)
series_clu = df.loc[k_1].values[:,1:]
n_clusters = 4
#plt.figure(figsize=(30,30))
sdtw_km = TimeSeriesKMeans(n_clusters, metric="softdtw",
metric_params={"gamma_sdtw": .01},
                           verbose=False, random_state=seed, dtw_inertia = True)
y_pred = sdtw_km.fit_predict(series_clu)

for yi in range(n_clusters):
    plt.subplot(n_clusters, 1, 1 + yi)
    for xx in series_clu[y_pred == yi]:
        plt.plot(xx.ravel(), "k-", alpha=.2)
    plt.plot(sdtw_km.cluster_centers_[yi].ravel(), "r-")
    #plt.xlim(0, sz)
    #plt.ylim(-4, 4)
    #if yi == 1:
        #plt.title("Soft-DTW $k$-means")

plt.tight_layout()
plt.show()
```

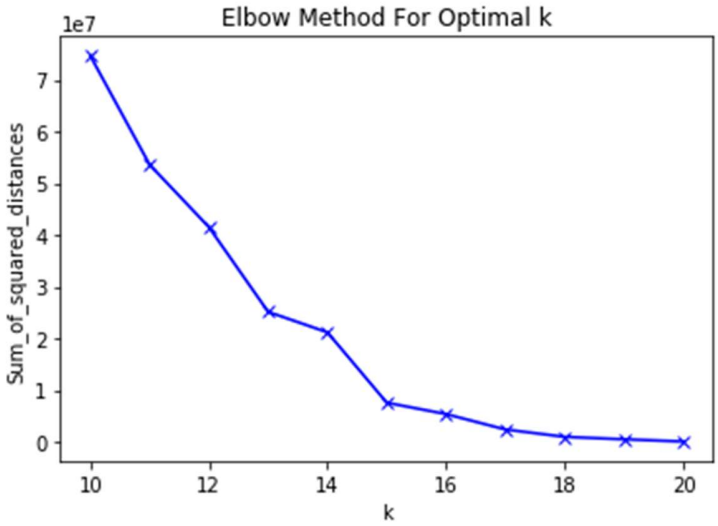
Output:



Optimitizing DTW:

```
series_clu = df_values
# Soft-DTW-k-means
print("Soft-DTW k-means")

K =  range(10,21)
Sum_of_squared_distances = []
#plt.figure(figsize=(30,30))
for k in K:
  n_clusters=k
  #sdtw_km = TimeSeriesKMeans(n_clusters, metric="softdtw", metric_params={"gamma_sdtw": .01},
  # verbose=False, random_state=seed, dtw_inertia = True).fit(series_clu)
  dtw_km = TimeSeriesKMeans(n_clusters, metric="dtw", max_iter=5, max_iter_barycenter=5, verbose=False, dtw_inertia = True, random_state = 0).fit(series_clu)
  Sum_of_squared_distances.append(dtw_km.inertia_)

plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

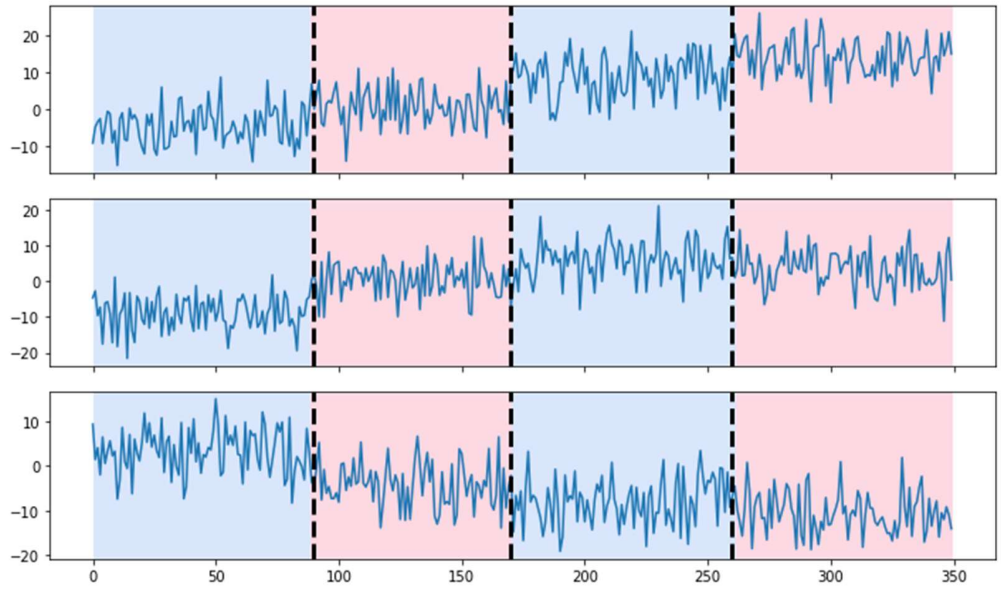
Output:



Event correlation:

```
n, dim = 350, 3  # number of samples, dimension
n_bkps, sigma = 3, 5  # number of change points, noise standart deviation
signal, bkps = rpt.pw_constant(n, dim, n_bkps, noise_std=sigma)
# change point detection
model = "l2"  # "l1", "rbf", "linear", "normal", "ar"
algo = rpt.Window(width=40, model=model).fit(signal)
my_bkps = algo.predict(n_bkps=3)

# show results
rpt.show.display(signal, bkps, my_bkps, figsize=(10, 6))
plt.show()
```

Output:



9. Cites

1.- https://www.bls.gov/cpi/

2.- https://en.wikipedia.org/wiki/Gross_domestic_product

3.- https://en.wikipedia.org/wiki/Gross_value_added

4.- https://www.investopedia.com/terms/i/interestrate.asp

5.- https://en.wikipedia.org/wiki/Principal_component_analysis

6.https://www.researchgate.net/publication/236015341_Interpretation_of_singular_spectrum_analysis_as_complete_eigenfilter_decomposition

7.- http://ssa.cf.ac.uk/zhigljavsky/pdfs/SSA/SSA_encyclopedia.pdf

8.- https://www.sciencedirect.com/science/article/pii/S0898122110003858

9.- http://www.mathcs.emory.edu/~lxiong/cs730_s13/share/slides/searching_sigkdd2012_DTW.pdf

10.- https://link.springer.com/content/pdf/10.1007/s10994-005-5828-3.pdf

11.-https://medium.com/@shachiakyaagba_41915/dynamic-time-warping-with-time-series-1f5c05fb8950

12.-https://s3.amazonaws.com/academia.edu.documents/45800856/Soheily-KhahPIATSA.pdf?response-content-disposition=inline%3B%20filename%3DProgressive_and_Iterative_Approaches_for.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20190702%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190702T081644Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=b2b4827faf2363b8d94d2516806cf8dcd70d62044b37c8d1d2d5780f9b7a94c7

13.- https://www.francois-petitjean.com/Research/Petitjean2014-ICDM-DTW.pdf

14 .- https://www.statisticshowto.datasciencecentral.com/autoregressive-model/

15.- https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q