MACHINE LEARNING FOR AFRICAN AIR FREIGHT

Boutaina Hajjar and Omar Drissi Kaitouni
Industrial Engineering, EMISYS Research Team, Mohammadia School of Engineering Mohammed V. University
Av. IbnSina, B. P. Agdal, Rabat, Morocco
E-Mail: Boutainahajjar@research.emi.ac.ma

ABSTRACT

Air freight transportation plays a pivotal role in stimulating economic development and enabling supply chains growth. Air cargo drives global trade by carrying more than 6.7 $ trillion of goods' value. Amid Covid Crisis, airfreight traffic remains essential in carrying nearly 1.5 million metric tons of medical equipment. Compared to normal demand pattern of 2019, air cargo volumes reached the levels of pre-pandemic period. In a dismal world air cargo market, African countries have continued to lead the international growth and reached by 2021 the highest pace of progress with 22.4% of growth. To sustain a positive pattern of development in Africa, it is relevant to develop a forecasting model using machine learning applications. Forecasting future market performance is essential for empowering planning processes. Nevertheless, only few researches have been developed for African countries. The main objective of this paper is to examine the driving factors of air cargo volume in Africa. For this purpose, we have applied machine learning algorithm to enhance the accuracy of data input and develop a reliable forecasting model. Findings and results emphasize that 28265774701 TY market of air cargo volume in Africa will contribute to GDP expansion by 13% over the next 20 years. This performance will help key stakeholders to improve African market and develop further prospects leading its potential growth.

Keywords: machine learning, forecasting, air cargo, data analysis.

INTRODUCTION

Air transportation industry is acting as a spur to global business and economic prosperity. Worldwide network of air transport facilitates world trade by providing greater access to international markets and improving the efficiency of business supply chain. The emergence of just-in-time (JIT) production operations and information technology, namely, e-commerce and e-business have fostered the growth of airfreight sector in the last two decades. During Covid period, air cargo remains essential to moving medical and pharmaceutical equipment and sustaining economies (Bartle, J. R., Lutte, R. K., & Leuenberger, D. Z., 2021). In fact, air cargo industry is playing a vital role in globalization. The development of international market at both the demand and the supply side have supported a steady growth of airfreight business. Despite representing only one percent by volume, international air cargo amounts to high ratio of value, accounting nearly 35 percent of world trade (Merkert, R., Van de Voorde, E., & de Wit, J., 2017). Though airfreight used to be considered as a "by-product" of passenger transport, dedicated freighter fleet has been developed to provide sufficient capacity to maximize profit. Air cargo volume is statistically linked to gross domestic Product (GDP) and GDP per capita measuring the economic development. Since 1970, the market has experienced a positive trend increase by doubling its volume every 10 years (Chang, Y. H., Yeh, C. H., & Wang, S. Y., 2007). As stated by (Kasarda, J. D., & Green, J. D., 2005), the growth of airfreight has outpaced trade expansion based on the World Bank analysis, as 132 percent of trade leads to 302 percent of airfreight increase for 68 countries between 1980 and 2000. After a severe downturn in 2008, global air cargo industry has rebounded since mid-2014 (Airplanes, B. C., 2018). Referring to Boeing's world air cargo forecast, this growth trend was projected to rise by 4.2 percent in the next 20 years reaching 584 billion Revenue Tonne-Kilometers (RTKs) in 2037. The International Air Transport Association (IATA) reported that this growth pattern was not identical for all regions based on airfreight analysis in 2017. While international Freight Tonne Kilometers (FTKs) has grown by 9 percent in 2017, posting its strongest performance since 2010, Africa has outperformed all regions by a large margin of 25.2 percent year on year FTKs.

African airfreight industry was integrated on an upward path into the world's market in recent years. Driven by foreign investments and an increasing trade flows with Asia, regional air cargo market has expanded its share by 64 percent on the route linking Africa to Asia according to IATA statistics of 2017. Furthermore, intra-African air cargo flows have increased by 8.2 percent, accounting approximately 152 000 Tonnes over 2017. Representing less than 2 percent of world share, airfreight in Africa has a promising potential of growth. Because of the highest number of remote countries, nearly 54 states, the continent is seeking to enhance air cargo traffic for an effective access to regional and international markets (Meichsner, N. A., O'Connell, J. F., & Warnock-Smith, D., 2018). Thus, Africa has an encouraging economic and trade performance, which are undoubtedly linked to airfreight efficiency. The International Civil Aviation Organization (ICAO) reported that the continent will outpace world's dynamics with a growth rate of 3.9 percent annually over 2045. Nevertheless, African airfreight market is struggling with sparse demand. This latter has dropped by 1.3 percent in 2018, showing the weakest ratio amid all regions based on IATA annual review 2019. The unbalanced demand and supply in Africa impedes the growth of air cargo market (Adler, N., Njoya, E. T., & Volta, N., 2018). Thereby, balancing the capacity of air cargo volume is essential to deal with future traffic. Consequently, developing a reliable forecasting model of African airfreight volume is a prominent topic to tackle.

With recent strides in artificial intelligence (AIs), machine learning (ML) methods have emerged to avoid spurious estimates of traditional approaches and produce accurate predictions of air cargo traffic. In fact, machine learning based modelling is quite robust to the problem of multicollinearity with which statistical techniques are compromised. These ML methods encompass three categories termed supervised learning, unsupervised learning and reinforcement learning. In supervised learning, the prediction model of an output is built using a known input and output data based on various models, namely linear regression, artificial neural networks (ANN), and k-nearest neighbors. In unsupervised learning, principal component analysis (PCA) and clustering methods are the most used to determine the hidden patterns of data. Reinforcement learning category comprises several techniques such as, Q-Learning to enable an agent to learn by interacting with its environment. Among all these methods, the widely used to predict air cargo was artificial neural networks (ANN) (Barua, L., Zou, B., & Zhou, Y., 2020). The first air cargo forecasting model based on ANN was developed to predict the demand from Japan to Taiwan by (Chen, Shu-Chuan, *et al*., 2012). In addition, (Loaiza, Miguel Figueroa, *et al.*, 2017; Baxter, G., & Srisaeng, P., 2018) have predicted airfreight demand for Columbian products and Australian exports. To enhance the performance of air cargo prediction, scholars have recommended hybrid models that combine machine learning and statistical techniques (e.g., linear modeling) (Sulistyowati, R., Kuswanto, H., & Astuti, E. T., 2018). Barua Zou and Zhou (2020) have underlined the great potential of machine learning as a stand-alone approach and have recommended the use of more diversified machine learning techniques for a predictive or descriptive analytics to come up with prominent results on air cargo forecasting.

Based on an extent of literature, most researchers have predicted air passenger because of the complexity of air cargo operations. Proposing a stable forecasting model helps airports and airlines to increase their capacities and provide to companies a reliable basis to plan infrastructure and potential demand related to transport technologies (Li, H., Bai, J., Cui, X., Li, Y., & Sun, S., 2020). Accurate air cargo forecasts is an essential component to enhance master plans on investment and service quality. In fact, airlines adopt various strategies to enhance air cargo operations. For instance, (Mongeau, M., & Bes, C., 2003) have highlighted the importance of aircraft loading optimization to address the problem of cargo loading. Moreover,(Totamane, R., Dasgupta, A., & Rao, S., 2012) have emphasized the main factors affecting cargo load factor such as, load balancing and efficient use of space. Furthermore, (Fok, K., & Chun, A., 2004) have developed a mathematical optimization to forecast cargo and mail load factors. As aforementioned, forecasting air cargo traffic helps airlines to perform efficiently in a competitive market. Since the major problem for airlines is capacity planning, studies on overbooking have attracted attention of most scholars (Klindokmai, Sirikhorn, *et al.*, 1014). Given that airfreight capacity is subject to significant volatility, cargo volumes and weights (capacity) are sold to freight forwarders twelve or six months in advance (Amaruchkul, K., Cooper, W. L., & Gupta, D., 2011). The uncertainty of air cargo market hampers researchers from balancing adequately the demand and supply of customers' requirements. Thus, developing a reliable forecasting assist airfreight operators to predict capacity demand. In fact, forecasting based on chargeable weight, which is considered as a key element of capacity planning in air cargo operations, leads to better estimates. Therefore, our study uses chargeable weight to predict African air cargo volume based on a quantitative forecasting technique. Although many studies on models evaluation were conducted, no specific forecasting approach was selected for all scenarios. The most common freight modelling methods were conducted under regression and time series (Farrington, P. A., 2011). However, this latter on their own are inefficient because of market fluctuations. Therefore, Barua Zou and Zhou (2020) have recommended to combine several forecasting approaches instead of applying a single method to reduce risks.

The models adopted to develop air cargo forecasting encompass four categories as argued by (Huang, K., & Lu, H., 2015): physical approaches, econometric models, hybrid (or combination) models and artificial intelligence (AI) methods. Physical approaches are not appropriate for short term forecasting as they are used to predict airfreight demand to plan an airport capacity extension. Econometric models investigate the correlation between external factors and air cargo demand (Wang, F., Zhuo, X., & Niu, B., 2017). However, these models are not reliable for a non-linear series. In recent decades, with the proliferation of computer techniques and intelligence learning, numerous studies have underlined the importance of artificial intelligence (AI). Machine learning is considered as a promising approach for air cargo demand forecasting to avoid, particularly, the problem of non-linearity in data. Some studies have investigated the effectiveness of AIs methods. (Zhang, G. P., & Qi, M., 2005) have underlined the importance of data processing in reducing errors. (Karlaftis, M. G., & Vlahogianni, E. I., 2011) have studied the suitable method for data analysis and highlighted its importance compared to tools used. According to Barua Zou and Zhou (2020), large and complex data require the use of machine learning based optimization which is especially pertinent for operations planning. Hybrid approaches have been also recommended by scholars to improve the accuracy and stability of forecasting. Hence, this paper uses multiple linear regression and machine learning models within principal component analysis (PCA) technique to overcome limitations and build an appropriate forecasting model.

Based on recent studies, (Hassan, L. A. H., Mahmassani, H. S., & Chen, Y., 2020) have recommended an effective demand forecasting model for freight operations based on combining time series and machine learning techniques in a reinforcement Learning context. Furthermore, Barua Zou and Zhou (2020) have underlined the importance of machine learning methods in predicting air cargo demand. In the African context, (Adenigbo, J. A., 2016) has applied multiple linear regression (MLR) to investigate the factors influencing agents choice to handle air cargo operations in Abuja airport located in Nigeria. Although forecasting studies have shown an increasing interest in recent years, no study has clearly developed a forecasting model for Africa to predict the future air cargo traffic and to investigate the factors underlying its growth. To sustain the applicability of various machine learning techniques, our study explores the use of principal component analysis (PCA) and linear regression, namely multiple linear regression analysis (MLR), to build a forecasting model of African air cargo volume and to underline the causal relationship with economic expansion. In fact, (Chang, Y. H., & Chang, Y. W., 2009) have demonstrated that there is a bi-directional relationship between airfreight expansion and economic development in the case of Taiwan. Our paper will examine the statistical relationship for Africa based on the predicted air cargo volume over the next 20 years. Therefore, we will analyze data input based on PCA method and identifies the key determinants of African air cargo traffic volume using MLR model. Our aim is to provide a valuable insights on African air cargo sector to help policymakers in planning processes. The remainder of this paper is organized as follows. Section 2 describes the methodology framework adopted for data collection and the processing of missing values. Section 3 details the use of machine learning to impute missing scores with PCA technique and to builda forecasting model with its parameters. Then, sections 4 and 5discuss the results and highlight the driving factors of African air cargo volume. Finally, we draw main conclusions and emphasize further suggestions for future research directions.

METHODOLOGY FRAMEWORK

Data Collection

In this paper, we are using real operational data collected from IATA CargoIS and IATA statistics 2017. Data sourced from IATA CargoIs show the volume of intra-regional air cargo imports and exports (air cargo volume between Africa, Europe, Asia Pacific, North Asia, Commonwealth of Independent States, Latin America and The Caribbean, Middle East and North Africa, North Atlantic and North America) expressed in “TY Mkt Weight” which stands for “This Year Market Weight (chargeable)” expressed in Kilograms (KG). Moreover, data collected from IATA air transport statistical results and IATA air freight market analysis define the statistics of the variables plotted in the forecasting model such as, Freight Tonne Kilometer (FTK), Revenue tonne kilometer (RTK), Weight Load Factor (WLF), and Freight Load Factor (FLF).

This section draws the overall approach as described in Figure-1 to forecast African air freight volume over the next 20 years based on machine learning techniques by combining Principal component analysis (PCA) and multiple linear regression (MLR). The aim was to develop a suitable forecasting model by 2037 with a combination of explanatory variables and assess the factors influencing air cargo volume in Africa.

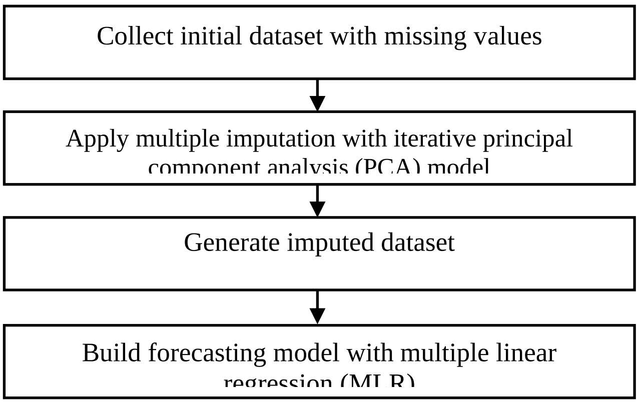

Collect initial dataset with missing values
Apply multiple imputation with iterative principal component analysis (PCA) model
Generate imputed dataset
Build forecasting model with multiple linear regression (MLR)

Figure-1. Steps of the proposed methodology.

MISSING DATA PROCESSING

Missing Data Description

Missing values are associated to the general term coarsened data that refers to a broad class of censored information (Schafer, J. L., & Graham, J. W., 2002). Unfortunately, the presence of missing scores in the collected dataset is an acute challenge for researchers mainly in scientific field. According to (Graham, J. W., 2009), missingness constitutes a drastic problem, particularly, for longitudinal studies implying multiple measurements on the same individuals. Hence, “the best thing to do with missing values is not to have any” (Josse, J., 2016). Three types of missing data were categorized by (Rubin, D. B., 1976): Missing Completely at Random (MCAR), missing at Random (MAR), and Missing Not at Random (MNAR). With MAR, the missingness (i.e., whether the data are missing or not) is related to the observed information, but unrelated to unobserved data. MCAR is a special case of MAR in which the probability of missingness does not depend on the observed data nor on the unknown missing data. Data are called MNAR, if the missingness does depend on unobserved data.

Historically, researchers have stated that one of the major reasons of missingness is non-response (Huisman, M., & Krause, R. W., 2018), and distinguished two types: unit of non-response, where the entire data are completely missing, and item of non-response where partial data are available. These types determine the amount of collected observations. In longitudinal studies, survey methodologists identified a third type of non-response called wave nonresponse which occurs when data are available at only some points as reported by Schafer and Graham in 2002.

Imputation of Missing Data

Survey statisticians handle often missing scores, especially in large datasets, by substituting plausible values for missing items; this practice is called imputation as stated by Schafer and Graham in 2002. As reported by Huisman and Krausein 2018 and defined by (Dempster, A., & Rubin, D., 1983) imputation is seductive and dangerous: "It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases". The main positive feature of imputation is to use the complete information on the observed variables to predict the missing values and to develop statistical analyses of the complete data. Until the 1970s, statisticians have dealt with missing values by editing as stated by Rubin's 1976. Traditionally, missing data were computed based on case deletion and single imputation approaches according to Schafer and Graham in 2002. Recently, more sophisticated procedures have been developed to underline the importance of multiple imputation (MI) methods.

The method of single or simple imputation encompasses replacing missing data by one estimate. For instance, replacing missingness by the mean score is termed mean imputation. Using regression model to impute missing data on the complete dataset is called regression-based single imputation or predicted mean imputation. According to Huisman and Krausein 2018, single imputation comprises four classes. The first class called imputing unconditional means is the most popular procedure in which missing data are replaced by the mean over the observed values. Although the averages are preserved, other parameter estimations such as variances and covariance's are distorted. Hence, survey statisticians have recommended the second group named imputing from unconditional distributions to reduce the bias of mean imputation. In this case, an approach named "hot deck" has emerged to preserve the distribution of variables by replacing missing values by a given value from the same dataset. Though this approach sustains means and variances, relations between variables still biased. Another class termed imputing conditional means helps to predict the mean scores of missing data using a formal models, such as linear regression to produce an accurate estimates especially under MAR assumption. Nevertheless, the approach was not recommended by Huisman and Krausein 2018 to analyze correlations because it underestimates variances and overstates covariances. Therefore, methodologists have underlined the usefulness of the last class termed imputing from conditional distributions to reduce the distortion generated from the previous group. This approach assumed that each missing value is substituted by a random draws from the conditional distribution. However, this method is quite complex and requires the properties of multiple imputation (MI) to generate a precise estimates. As reported by Huisman and Krausein 2018, single imputation is effective if analyses do not require confidence intervals or assumptions testing. Single imputation overestimates the precision of inferences and understates the standard errors. Thereby, this method produces potentially biased outcomes and reflects the issue of under-coverage related to the uncertainty of missingness. As reported by Schafer and Graham in 2002 this problem is solved using multiple imputations, which is recommended as the best procedure, especially for medium to large amounts of missing values.

In modern imputation procedures, multiple imputation (MI) has been highly recommended by statisticians as the most promising method to deal with missing values. Recognized as a standard approach for various areas of research, MI results in a wider confidence intervals and more accurate standard errors leading to unbiased estimates in most statistical software. Though, this procedure generates precise results under MAR assumption, it builds as well on both MCAR and MNAR. In MI, each missing data is substituted multiple times (i.e., m times) with plausible values based on the distribution of other scores in the dataset. The imputed estimates will not only affect the missing values, but also further items of the observed data. For precise estimates, many researchers have investigated the number of imputed datasets. Recent studies have proposed the rule of thumb where the percentage of missingness is equal to the number of imputations (Enders, C. K., 2017). Moreover, a compromised solution has been underlined by (Pedersen, Alma B., *et al.*, 2017), suggesting three to five imputed datasets for model building. The imputation model as highlighted by Enders in 2017 requires the introduction of additional variables to enhance the plausibility of MAR assumption, called auxiliary variables. These variables, identified using simple bivariate correlations, have been suggested by statisticians to improve the power of estimates. Multiple imputation implies three different stages according to Enders in 2017: imputing, analyzing, and combining as illustrated in Figure-2 developed by Pedersen, Alma B., *et al*. in 2017. The first phase is based on creating imputations with an iterative algorithm and updating regression parameters of the filled data for each next imputation using Bayesian estimation procedures. In Bayesian theory researchers combine preceding distribution of parameters and a likelihood function to allow a good performance of multiply imputed datasets (Chen, Qixuan, *et al*., 2018). Subsequently, statistical analyses are performed to create parameter estimates and standard errors for each imputed dataset. At the pooling phase, a set of estimates and standard errors are aggregated using software packages for procedures simplification. Computationally, several packages such as SPSS and R are used to facilitate the implementation of the model building.

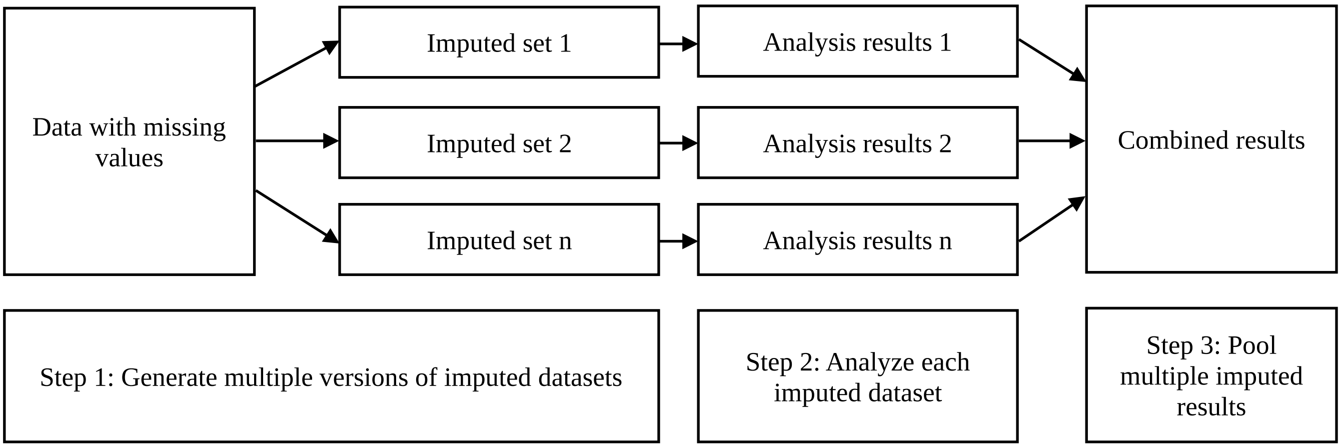


Figure-2. Phases of multiple imputation procedure (Pedersen, Alma B., *et al*., 2017)

Multiple imputation solves the problem of understating standard errors and yields good coverage rate. In general, MI considers the uncertainty in the estimate using either a Bayesian or bootstrap approach (Audigier, V., Husson, F., & Josse, J., 2016). This latter consists in generating multiple incomplete datasets and estimating parameters for each bootstrap replication. In Bayesian method parameters are derived from the subsequent distribution of the model based on the prior distribution and the observed data. To perform multiple imputation, researchers have proposed two classical models namely the explicit joint model and chained equations (i.e., conditional modelling). While joint model assumes assumption of normality on variables, chained equations appears more flexible since it tailored a model for each variable with missingness, and variables are consecutively imputed using these models. However, both classical models cannot perform properly where the number of variables exceeds the number of individuals or in case of high correlation of variables. Hence, new flexible methods of multiple imputation based on principal component analysis have been emerged recently to enhance the quality of imputation.

MACHINE LEARNING ANALYSIS

Multiple Imputation with Principal Component Analysis

Firstly, we need to prepare data for linear regression. Therefore, it is relevant to ensure Data cleaning by imputing missing scores, eliminating errors and ensuring no correlation between independent variables. In our study, we propose multiple imputation method based on iterative PCA to handle missing data. The results are obtained using R package missMDA.

For a proper statistical analysis, principal components (PCs) methods have been recommended to describe and visualize the observed data. Given the nature of variables, these methods include numerous approaches such as principal component analysis (PCA) which is an unsupervised learning technique as explained previously. As defined by (Josse, J., Pagès, J., & Husson, F., 2011), PCA is a multivariate statistical tool used to reduce a high dimensionality of data without reducing the data variability for a better interpretation of variables. Several algorithms have been suggested, including the iterative PCA algorithm detailed by (Josse, J., & Husson, F., 2012). When missing values, the algorithm implies three main steps. First, missing data are substituted with an initial values. Subsequently, PCA is applied on the imputed table with S dimensions and missing data are imputed using PCA. Then, initial values and standard deviations of each variable are updated. Finally, iterations are computed until convergence. With a large set of missing data, iterative PCA suffers from overfitting problem, which is solved by regularization. In practice, the R package “missMDA” represents a regularized version of the algorithm. In addition, Josse, Pagès and Husson, have proposed in 2011 multiple imputation method based on PCA to preserve the variability of parameters for each imputation due to missing values. In general, researchers have recommended the use of PCA as a tool of multiple imputation to obtain valid variances and estimates. In fact, Josse has conducted a simulation study to assess the method of multiple imputation with PCA model. While the classical approaches, namely joint and conditional modelling underperform when regressions have difficulties, multiple imputation using principal component analysis for continuous variables provides valid results with a good coverage (i.e., percentage of confidence intervals including the parameter value). Indeed, the imputation of missing values based on PCA, has emerged as a flexible tool since it can be applied on large or small datasets with weak or strong correlations between variables or with the number of variables less or greater than the number of individuals. Moreover, the method remains competitive in case of linearity between variables, and particularly when generating variables from a PCA model.

Multiple Linear Regression

Linear regression is a modelling technique for data analysis aiming to make effective predictions (Tranmer, M., & Elliot, M., 2008). Though simple linear regression is established on bivariate model to predict a dependent variable from an explanatory variable, multiple linear regression (MLR) produces a multivariate model including more than a single explanatory variable. In our study, we will build a prediction model under MLR method using IBM program SPSS statistics version 26 to investigate the relation between our dependent variable Y_{i} which is African air cargo volume and the combination of explanatory variables X_i(air cargo volume betweenEurope, Asia Pacific, North Asia, Commonwealth of Independent States, Latin America and The Caribbean, Middle East and North Africa, North Atlantic and North America, % Weight load factor, % Year-on-year change in Freight load factor (% FLF), Freight tonne-km (millions),% in world traffic of freight tonne-km, Revenue tonne -km (millions), % in world traffic of Revenue tonne -km, Tonne kilometers available (millions), % in world traffic of Tonne kilometers available, % of industry FTKs in 2017 (World share), % Year-on-year of Freight Load factor level (FLF Level)).

The equation of multiple liner regression is written under the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Where k is the number of independent variables,β_0 is the constant, $\beta_1, \beta_2 \ldots \beta_k$ are the amount of Y_i increase for an increase in one value of X_i, and ε_i is the error term (i.e., the difference between predicted and actual values). The building of our multiple linear regression is described in Figure-3.

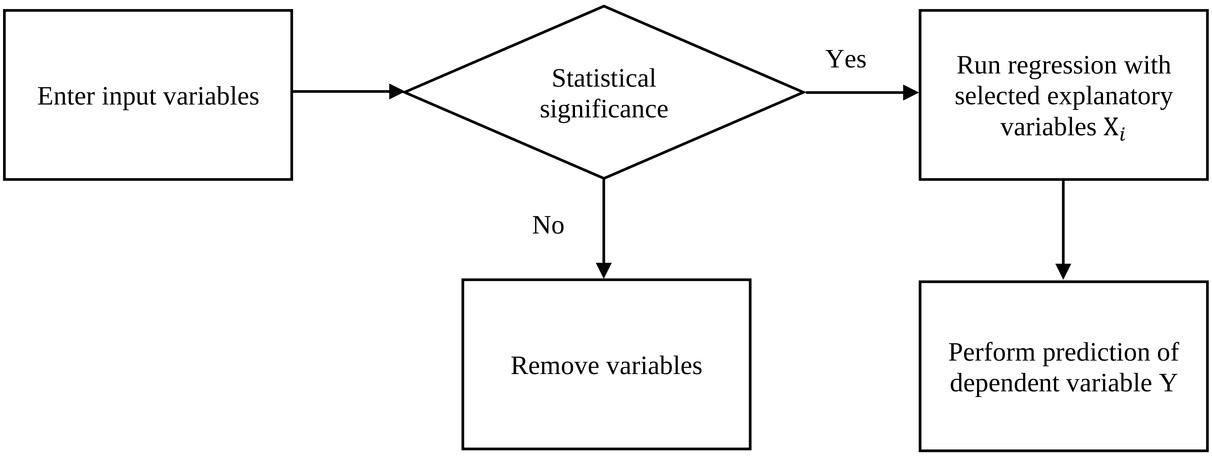

Enter input variables
Statistical significance
Yes
No
Run regression with selected explanatory variables X_i
Remove variables
Perform prediction of dependent variable Y

Figure-3. Multiple linear regression building.

Regression outputs imply model summary, ANOVA and Coefficients tables. The results of our study are discussed and interpreted in results and discussions section.

The model summary provides the results of “R, R², adjusted R²and the standard error of the estimate” to determine how well the regression model fits the data. In fact, R represents the multiple correlation coefficient and measures the quality of prediction of the outcome (i.e., dependent variable). The multiple coefficient of determination or R² (R-squared) value describes the proportion of variance in the dependent variable which is explained by independent (i.e., predictor) variables. In general, R² ranges between 0 and 1. If the dependent value is well described by the explanatory variables, then R² will take a value close to 1. However, R² is close to zero in case of poor linear predictor to Y. R² must be combined with other residual plots because a good model can have a low R² value and a biased model can have a high R² value as demonstrated by (Harel, O., 2009). Therefore, adjusted R² (adj. R²) is another important factor used for better interpretation and embodied a modified version of R² for the number of predictors in the model. High difference between R² and adjusted R² implies a bad fit of the regression model. Adjusted R² is always less than R². In addition, the standard error of the estimate or standard deviation of the residuals shows the spread of the distribution to measure the precision of regression model. Whereas multicollinearity alters the model, the smallest standard error is the most required for better predictions of variables estimates.

ANOVA table is generated as a regression output to analyze the variance. In fact, “F-ratio” plotted in the table tests whether the data fit the regression model. The Sig. value commonly known as “p-value” demonstrates the statistical significance of a value. The regression model shows a good fit and is statistically significant when “p-value” is below 0.05.

Statistical significance of explanatory variables is analyzed in the coefficients table, which is another output of regression, where just essential predictors that better explain the dependent variable are displayed. As we will describe later in results and discussions section, B column shows β parameters for the predicted equation. Standard errors for coefficients as mentioned afore should have the smallest values. The Sig. or p-value describes whether the estimates show a statistical significance different from zero. Though unstandardized coefficients explain the variation of predictors with an outcome when all other independent variables are constant, standardized parameters termed Beta weights explain the increases of the outcome for one standard deviation increase of a predictor.

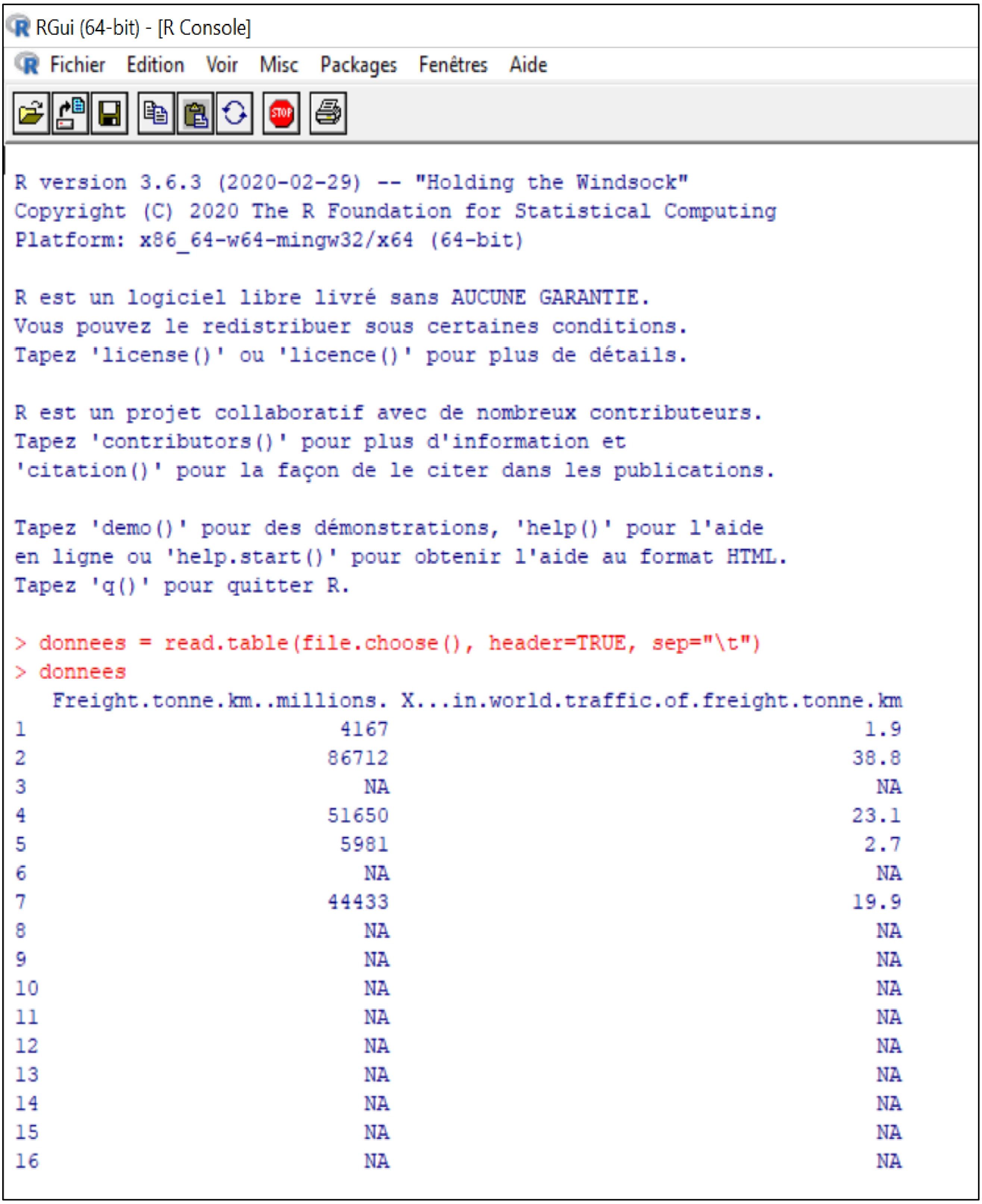
RESULTS AND DISCUSSIONS

Data Imputation in PCA Method

The tool used to impute missing values of our data matrix based on iterative principal component analysis (PCA) model is R version 3.6.3. On modern personal computer. We have implemented multiple imputation based on iterative PCA model to obtain an accurate imputed dataset with good prediction of missing scores. The R package miss MDA used by default for the regularized iterative PCA algorithm was adopted to solve the problem of matrix completion. This popular algorithm in machine learning community improves efficiency and reduces bias of inferences. Moreover, we have applied multiple imputation to reflect the variability in missing scores before analyzing statistically the variable of our interest which is air cargo volume in Africa. This method has been recommended for its good coverage and results accuracy.

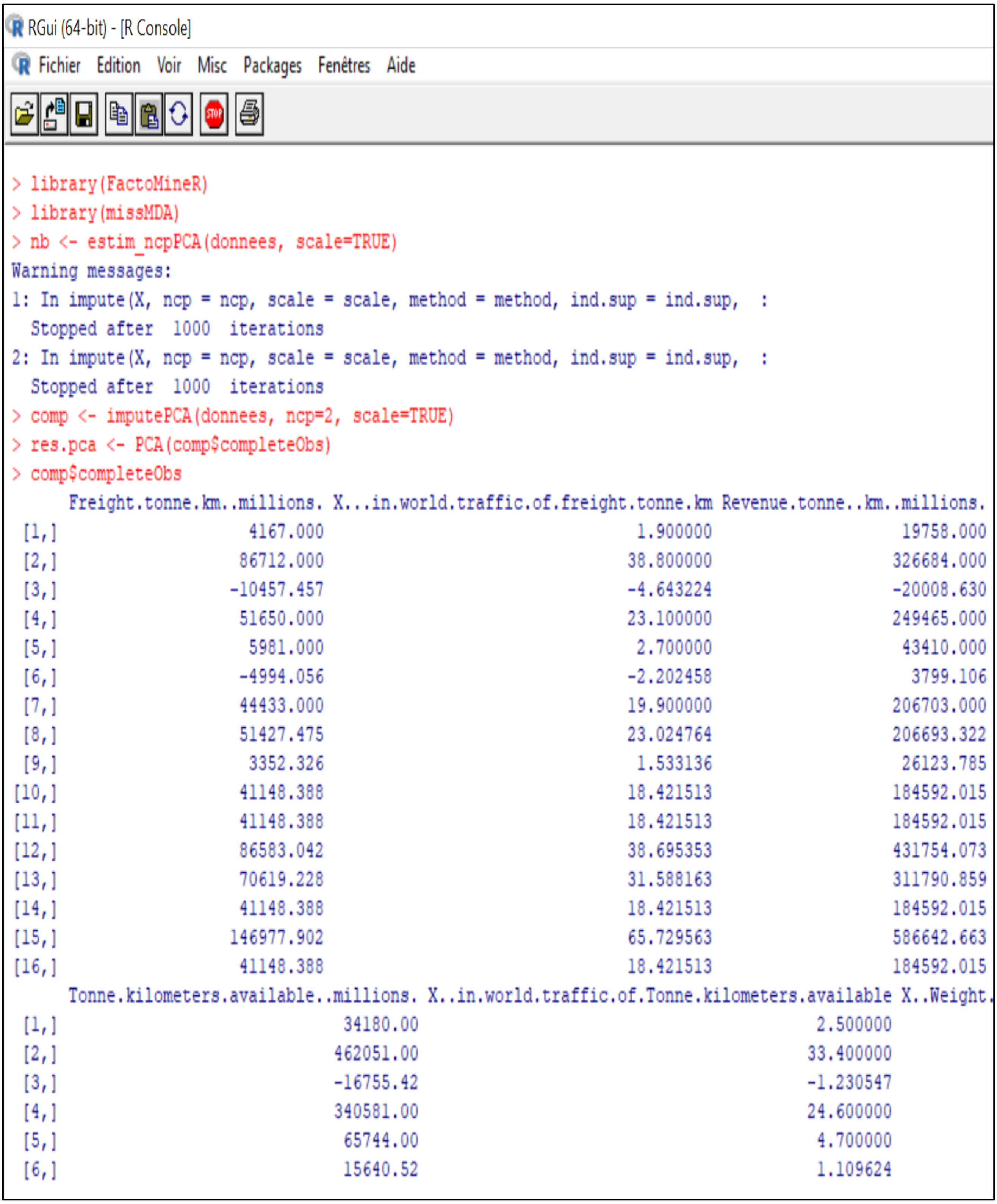
After loading raw data with R language as shown in the R code of figure 4, we have followed the code lines shown in Figure-5. Then, we have visualized the pattern of missing data with the R package FactoMineR. Afterward, we have defined the number of components to be used for the reconstruction formula in the function estim_ncpPCA. Subsequently, we have used impute PCA to perform principal components analysis method for our incomplete dataset and to impute data with principal components methods. At convergence and after reconstructing data with corresponding ncp (e.g., ncp=2 which is in our case the optimal number of dimensions when imputing missing data), we have performed PCA on the complete dataset and plotted uncertainties as shown in figure 6 through the function res.pca. Hence, the final imputed matrix is obtained in the object complete Obs with an estimation of scores.

We have implemented multiple imputation based on iterative PCA model to obtain an accurate imputed dataset with good prediction of missing scores. The R package miss MDA used by default for the regularized iterative PCA algorithm was adopted to solve the problem of matrix completion. This popular algorithm in machine learning community improves efficiency and reduces bias of inferences. Moreover, we have applied multiple imputation to reflect the variability in missing scores before analyzing statistically the variable of our interest. This method has been recommended for its good coverage and results accuracy.


RGui (64-bit) - [R Console]
Fichier Edition Voir Misc Packages Fenêtres Aide

```
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

> donnees = read.table(file.choose(), header=TRUE, sep="\t")
> donnees
   Freight.tonne.km..millions. X...in.world.traffic.of.freight.tonne.km
1                         4167                                      1.9
2                        86712                                     38.8
3                           NA                                       NA
4                        51650                                     23.1
5                         5981                                      2.7
6                           NA                                       NA
7                        44433                                     19.9
8                           NA                                       NA
9                           NA                                       NA
10                          NA                                       NA
11                          NA                                       NA
12                          NA                                       NA
13                          NA                                       NA
14                          NA                                       NA
15                          NA                                       NA
16                          NA                                       NA
```

Figure-4. Extract of raw data with missing scores loaded in R language.



```
> library(FactoMineR)
> library(missMDA)
> nb <- estim_ncpPCA(donnees, scale=TRUE)
Warning messages:
1: In impute(X, ncp = ncp, scale = scale, method = method, ind.sup = ind.sup,  :
  Stopped after  1000  iterations
2: In impute(X, ncp = ncp, scale = scale, method = method, ind.sup = ind.sup,  :
  Stopped after  1000  iterations
> comp <- imputePCA(donnees, ncp=2, scale=TRUE)
> res.pca <- PCA(comp$completeObs)
> comp$completeObs
      Freight.tonne.km..millions. X...in.world.traffic.of.freight.tonne.km Revenue.tonne..km..millions.
 [1,]                    4167.000                                 1.900000                    19758.000
 [2,]                   86712.000                                38.800000                   326684.000
 [3,]                  -10457.457                                -4.643224                   -20008.630
 [4,]                   51650.000                                23.100000                   249465.000
 [5,]                    5981.000                                 2.700000                    43410.000
 [6,]                   -4994.056                                -2.202458                     3799.106
 [7,]                   44433.000                                19.900000                   206703.000
 [8,]                   51427.475                                23.024764                   206693.322
 [9,]                    3352.326                                 1.533136                    26123.785
[10,]                   41148.388                                18.421513                   184592.015
[11,]                   41148.388                                18.421513                   184592.015
[12,]                   86583.042                                38.695353                   431754.073
[13,]                   70619.228                                31.588163                   311790.859
[14,]                   41148.388                                18.421513                   184592.015
[15,]                  146977.902                                65.729563                   586642.663
[16,]                   41148.388                                18.421513                   184592.015
      Tonne.kilometers.available..millions. X..in.world.traffic.of.Tonne.kilometers.available X..Weight.
 [1,]                              34180.00                                          2.500000
 [2,]                             462051.00                                         33.400000
 [3,]                             -16755.42                                         -1.230547
 [4,]                             340581.00                                         24.600000
 [5,]                              65744.00                                          4.700000
 [6,]                              15640.52                                          1.109624
```

Figure-5. Extract of imputed data with R package missMDA.

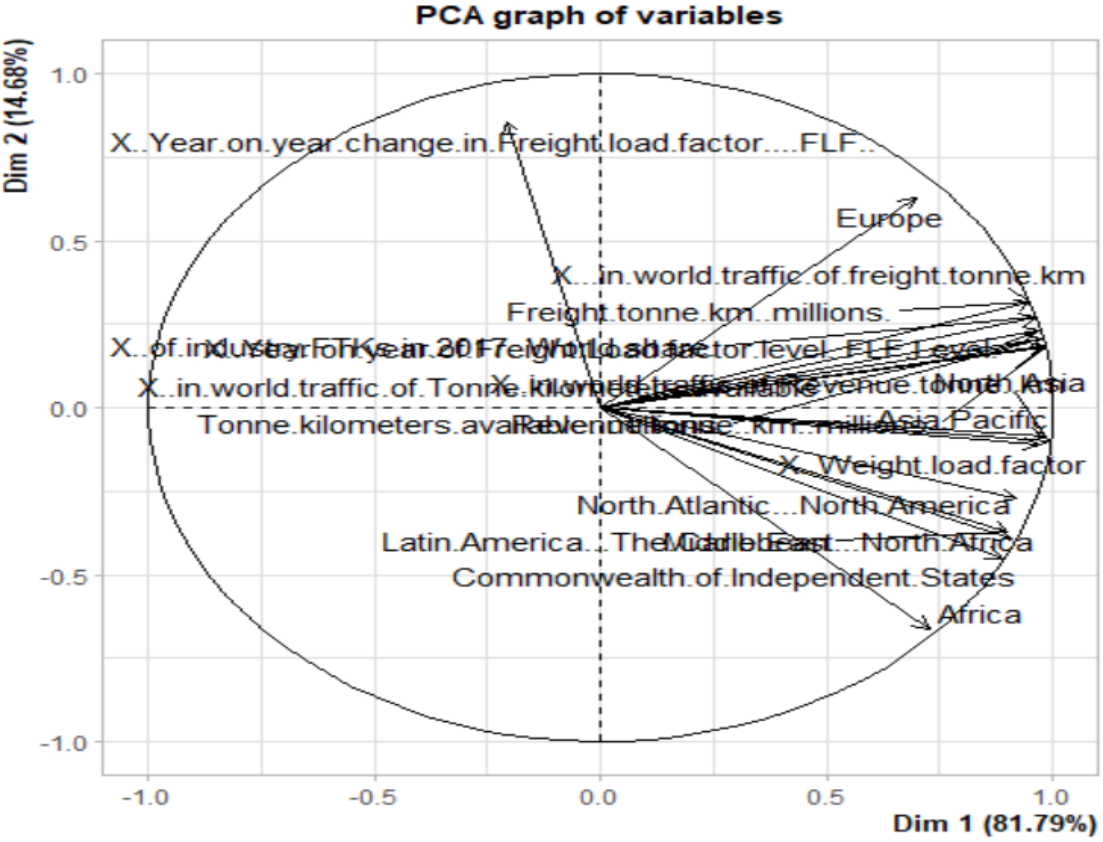


Figure-6. Observed PCA variables.

Building of Forecasting Model

As mentioned before, the tool used to perform our model is IBM SPSS Statistics version 26 on modern personal computer. Through SPSS statistics, we have selected standard regression termed forced entry which is the most widely used regression method where all independent variables are entered into the equation. In fact, our complete dataset has been analyzed using linear regression to investigate the relationship between our dependent variable Y which is air cargo volume in Africa, and the given explanatory variables X_i (air cargo volume between Europe, Asia Pacific, North Asia, Commonwealth of Independent States, Latin America and The Caribbean, Middle East and North Africa, North Atlantic and North America, % Weight load factor, % Year-on-year change in Freight load factor (% FLF), Freight tonne-km (millions),% in world traffic of freight tonne-km, Revenue tonne -km (millions), % in world traffic of Revenue tonne -km, Tonne kilometers available (millions), % in world traffic of Tonne kilometers available, % of industry FTKs in 2017 (World share), % Year-on-year of Freight Load factor level (FLF Level)) in the model.

Based on regression outputs, results of Table-1 determine how well our model fits the data. The value of R is 0.993 and it shows the quality of prediction of our dependent variable. The value of R² is 0.986 and it indicates that there are 98.6 % variation in African air cargo volume explained by predictors. Since the values of R and R² are close to 1, our regression equation is very useful for a good level of prediction. The value of adjusted R² is 0.964 and it determines that 96.4% of the variance in the dependent variable is explained by predictors which are to keep in the model. The small discrepancy between R² and adjusted R² confirms the goodness of fit of our model as ideally the multiple coefficient of determination should be very close to adjusted R². The standard error of the estimate represents the average distance between the observed values and the regression line to show the spread of the distribution. In our study, the estimates may be wrong by nearly 32824547, 07 TY Mkt Weight.

By considering the regression results and the afore-mentioned interpretation, our model demonstrates a best fit of the data.

Table-1. Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,993 a	,986	,964	32824547,07

a. Predictors: (Constant), air cargo volume in North Asia, % Year-on-year change in Freight load factor (% FLF), air cargo volume in Europe, air cargo volume in North Atlantic and North America, air cargo volume in Latin America and The Caribbean, air cargo volume in Middle East and North Africa, air cargo volume in Commonwealth of Independent States, % Weight load factor, air cargo volume in Asia Pacific

The statistical significance of our model based on F-ratio is explained through Table-2. In our regression model, explanatory variables statistically significantly predict African air cargo volume as F (9, 6) =45,747, p (.000)< 0.05. Hence, our regression model shows a good fit of the data.

Table-2. ANOVA^a^.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,436E+17	9	4,929E+16	45,747	,000b
	Residual	4,465E+15	6	1,077E+15		
	Total	4,501E+17	15			

a. Dependent Variable: Air cargo volume in Africa
b. Predictors: (Constant), air cargo volume in North Asia, % Year-on-year change in Freight load factor (% FLF), air cargo volume in Europe, air cargo volume in North Atlantic and North America, air cargo volume in Latin America and The Caribbean, air cargo volume in Middle East and North Africa, air cargo volume in Commonwealth of Independent States, % Weight load factor, air cargo volume in Asia Pacific

Since our model demonstrates a good fit of data, it is appropriate to build a regression equation based on the coefficients or estimated model parameters illustrated in Table-3. This coefficients table describes the statistical significance for each independent variable and identifies the variables to keep in the model by considering other predictors.

Table-3. Coefficients^a^.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1366645085	ε=3753389985		,364	,728
	X_1 :% Weight load factor	-18199223	ε_1=54397516,60	-,757	-,335	,749
	X_2 : % Year-on-year change in Freight load factor (% FLF)	-111696907	ε_2=209558303,7	-,254	-,533	,613
	X_3 : Air cargo volume inAsia Pacific	-,478	ε_3= 0,484	-2,426	-,987	,362
	X_4 : Air cargo volume inCommonwealth of Independent States	5,116	ε_4= 9,631	1,579	,531	,614
	X_5 : Air cargo volume inEurope	,189	ε_5= 0,172	,452	1,099	,314
	X_6 : Air cargo volume inLatin America and The Caribbean	-,477	ε_6 =0,216	-,786	-2,212	,069
	X_7: Air cargo volume inMiddle East and North Africa	,328	ε_7= 0,343	,672	,957	,375
	X_8: Air cargo volume inNorth Atlantic and North America	,225	ε_8 =0,627	,918	,359	,732
	X_9 : Air cargo volume inNorth Asia	,365	ε_9 =0,267	1,373	1,368	,220
a. Dependent Variable: Air cargo volume in Africa						

B column provides β coefficients for the regression equation. The positive parameters indicate that any increase of a predictor variable leads to an increase in air cargo volume in Africa. Contrarily, any increase of a predictor with negative coefficients cause a decrease of dependent variable. The constant 1366645085 called the interceptβ_0, is the explanatory value for the outcome variable air cargo volume in Africa if all independent variables have a null value.

Considering that multiple regression equation is written, as explained afore, under the form:

$$\mathrm{Y}_i = \beta_0 + \beta_1\chi_{1i} + \beta_2\chi_{2i} + \cdots + \beta_k\chi_{ki} + \varepsilon_i$$

Our regression model is defined as at the following equation:

$$\text{Air cargo volume in Africa} = \beta_0 + \beta_1 \times X_1 + \varepsilon_1 + \beta_2 \times X_2 + \varepsilon_2 + \beta_3 \times X_3 + \varepsilon_3 + \beta_4 \times X_4 + \varepsilon_4 + \beta_5 \times X_5 + \varepsilon_5 + \beta_6 \times X_6 + \varepsilon_6 + \beta_7 \times X_7 + \varepsilon_7 + \beta_8 \times X_8 + \varepsilon_8 + \beta_9 \times X_9 + \varepsilon_9 + \varepsilon$$

For the percentage of weight load factor (β=-18199223), any increase in percentage of X_1: weight load factor by one unit leads to a decrease of African air cargo volume by 18199223TY market by considering that the other predictors are constant. For the percentage of X_2 :year-on-year change in freight load factor (β=-111696907), any increase in percentage of weight load factor by one unit leads to a decrease of African air cargo volume by 111696907 TY market by considering that the other predictors are constant. For each one-unit increase in X_3:Asia pacific (β=-0,478), air cargo volume in Africa decreases by 0,478TY market by considering that the other predictors are constant. For each one-unit increase in X_4: Commonwealth of Independent States (β=5,116), air cargo volume in Africa increases by 5,116 TY market by considering that the other predictors are constant. For each one-unit increase in air cargo volume in X_5: Europe (β=0,189), air cargo volume in Africa increases by 0,189 TY market by considering that the other predictors are constant. For each one-unit increase in air cargo volume in X_6: Latin America and the Caribbean (β=-0,477), air cargo volume in Africa decreases by 0,477 TY market by considering that the other predictors are constant. For each one-unit increase in air cargo volume in X_7: Middle East and North Africa (β=0,328), air cargo volume in Africa increases by 0,328 TY market by considering that the other predictors are constant. For each one-unit increase in X_8:air cargo volume in North Atlantic and North America (β=0,225), air cargo volume in Africa increases by 0,225 TY market by considering that the other predictors are constant. For each one-unit increase in air cargo volume in X_9: North Asia (β=0,365), air cargo volume in Africa increases by 0,365 TY market by considering that the other predictors are constant.

By substituting the values of β coefficients from table 3, our prediction equation is written as below:

Air cargo volume in Africa = 1366645085 + (-18199223)×(% Weight load factor) + 54397516,60 + (-111696907)× (% Year-on-year change in Freight load factor (% FLF))+ 209558303,7+ (-0,478)×(air cargo volume in Asia pacific) +0,484 +5,116 × (Air cargo volume in Commonwealth of Independent States)+ 9,631 + 0,189× (Air cargo volume in Europe)+ 0,172+(-0,477) × (Air cargo volume in Latin America and The Caribbean)+ 0,216 + 0,328× (Air cargo volume in Middle East and North Africa)+ 0,343+ 0,225 × (Air cargo volume in North Atlantic and North America)+ 0,627+ 0,365 × (Air cargo volume in North Asia)+ 0,267+3753389985

The interpretation above was valid under no multicollinearity assumption. This condition is verified through correlation and collinearity statistics. As a rule of thumb, all the variables with tolerance <0.1 are excluded from the model as shown in Table-4.

Table-4. Excluded variables[a].

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Freight tonne-km (millions)	42,410[b]	4,086	,009	,877	6,145E-6
	% in world traffic of freight tonne-km	42,378[b]	4,086	,009	,877	6,155E-6
	Revenue tonne -km (millions)	45,649[b]	4,086	,009	,877	5,304E-6
	% in world traffic of Revenue tonne -km	45,611[b]	4,086	,009	,877	5,313E-6
	Tonne kilometers available (millions)	41,575[b]	4,086	,009	,877	6,395E-6
	% in world traffic of Tonne kilometers available	41,551[b]	4,086	,009	,877	6,402E-6
	% of industry FTKs in 2017 (World share)	46,469[b]	4,087	,009	,877	5,119E-6
	% Year-on-year of Freight Load factor level (FLF Level)	55,849[b]	4,087	,009	,877	3,544E-6

a. Dependent Variable: Air cargo volume in Africa
b. Predictors in the Model: (Constant), air cargo volume in North Asia, % Year-on-year change in Freight load factor (% FLF), air cargo volume in Europe, air cargo volume in North Atlantic and North America, air cargo volume in Latin America and The Caribbean, air cargo volume in Middle East and North Africa, air cargo volume in Commonwealth of Independent States, % Weight load factor, air cargo volume in Asia Pacific

RESULTS ANALYSIS

Based on our forecasting model, we have built a statistical relationship between African air cargo and the factors influencing its expansion over the next 20 years. The results show that the volume of airfreight at a regional level will reach nearly 28265774701 TY market by 2037. This outcome indicates that the annual growth rate in Africa will be estimated at 20% which represents 5 times the world annual growth rate of 4.2% by 2037. Furthermore, this paper shows the statistical relationship between GDP and air cargo expansion. In fact, African GDP will raise by 13% over the next 20 years compared to 3.3% estimated by Boeing, which represents 4 times increase of GDP in Africa. These results underline the importance of airfreight development on economic prosperity of African countries.

CONCLUSION AND RECOMMENDATION

This research provides an insight into a new forecasting approach application and investigates the factors influencing African air cargo volume. Considering the rapid development of machine learning techniques to predict air cargo traffic and their capability to harness the power of collected data, we have developed a new attempt combining machine learning algorithm termed principal component analysis and multiple linear regression to forecast air cargo volume in Africa.

By taking full advantage of machine learning procedures and traditional linear regression, the proposed forecasting method has been demonstrated to be useful in tightening prediction errors and enhancing air cargo forecasting accuracy. The results of our new methodology have been validated, since our regression model has shown a good fit of data and predicted the African air freight with a high level of prediction (R^2=98.6%). The model has revealed the relationship between independent variables to help policymakers in understanding and developing the potential capacity of African market. Moreover, the results indicate the importance of air cargo expansion on economic growth, as GDP will raise by 13% over the next 20 years for a volume of 28265774701 TY market.

In the future works, better forecasts can be drawn if data are extended in time. In addition, forecasting efficiency could be improved by utilizing further machine learning models with a large set of explanatory variables. This study constitutes a step forward in the domain of air freight forecasting and we recommend using the results as an input to more complex models of metaheuristic optimization algorithms. Further efforts should be deployed to tackle African air cargo topic and bring the best from the existing forecasting models.

ACKNOWLEDGMENT

We acknowledge The IATA CargoIs "The International Air Transport Association CargoIs", which is the air cargo industry's premier market intelligence tool, for supporting us by real data base.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

Bartle J. R., Lutte R. K. & Leuenberger D. Z. 2021. Sustainability and air freight transportation: Lessons from the global pandemic. Sustainability. 13(7): 3738.

Merkert R., Van de Voorde E. & de Wit J. 2017. Making or breaking-Key success factors in the air cargo market.

Chang Y. H., Yeh C. H. & Wang S. Y. 2007. A survey and optimization-based evaluation of development strategies for the air cargo industry. International Journal of Production Economics. 106(2): 550-562.

Kasarda J. D. & Green J. D. 2005. Air cargo as an economic development engine: A note on opportunities and constraints. Journal of Air Transport Management. 11(6): 459-462.

Airplanes B. C. 2018. World Air Cargo Forecast 2018-2037. URL: http://www. boeing. com.

Meichsner N. A., O'Connell J. F. & Warnock-Smith D. 2018. The future for African air transport: Learning from Ethiopian Airlines. Journal of Transport Geography. 71, 182-197.

Adler N., Njoya E. T. & Volta N. 2018. The multi-airline p-hub median problem applied to the African aviation market. Transportation Research Part A: Policy and Practice. 107, 187-202.

Barua L., Zou B. & Zhou Y. 2020. Machine learning for international freight transportation management: a comprehensive review. Research in Transportation Business & Management. 34, 100453.

Chen S. C., Kuo S. Y., Chang K. W. & Wang Y. T. 2012. Improving the forecasting accuracy of air passenger and air cargo demand: the application of back-propagation neural networks. Transportation Planning and Technology. 35(3): 373-392.

Loaiza M. F., Solano R. P., Simancas R. & Ojito V. H. 2017, April. Modeling Demand for Air Cargo in the Colombian Context. In 2017 International Conference on Advanced Materials Science and Civil Engineering (AMSCE 2017). Atlantis Press.

Baxter G. & Srisaeng P. 2018. The use of an artificial neural network to predict Australia’s export air Cargo demand. International Journal for Traffic and Transport Engineering. 8(1): 15-30.

Sulistyowati R., Kuswanto H. & Astuti E. T. 2018, March. Hybrid forecasting model to predict air passenger and cargo in Indonesia. In 2018 international conference on information and communications technology (ICOIACT) (pp. 442-447). IEEE.

Li H., Bai J., Cui X., Li Y. & Sun S. 2020. A new secondary decomposition-ensemble approach with cuckoo search optimization for air cargo forecasting. Applied Soft Computing. 90, 106161.

Mongeau M. & Bes C. 2003. Optimization of aircraft container loading. IEEE Transactions on aerospace and electronic systems. 39(1): 140-150.

Totamane R., Dasgupta A. & Rao S. 2012. Air cargo demand modeling and prediction. IEEE Systems Journal. 8(1): 52-62.

Fok K. & Chun A. 2004, August. Optimizing air cargo load planning and analysis. In Proceedings of the international conference on computing, communications and control technologies.

Klindokmai S., Neech P., Wu Y., Ojiako U., Chipulu M. & Marshall A. 2014. Evaluation of forecasting models for air cargo. The International Journal of Logistics Management.

Amaruchkul K., Cooper, W. L. & Gupta D. 2011. A note on air-cargo capacity contracts. Production and Operations Management. 20(1): 152-162.

Farrington P. A. 2011. Methods for forecasting freight in uncertainty: time series analysis of multiple factors (No. 930-768). University of Alabama.

Huang K. & Lu H. 2015. A linear programming-based method for the network revenue management problem of air cargo. Transportation Research Part C: Emerging Technologies. 59, 248-259.

Wang F., Zhuo X. & Niu B. 2017. Strategic entry to regional air cargo market under joint competition of demand and promised delivery time. Transportation Research Part B: Methodological. 104, 317-336.

Zhang G. P. & Qi, M. 2005. Neural network forecasting for seasonal and trend time series. European journal of operational research. 160(2): 501-514.

Karlaftis M. G. & Vlahogianni E. I. 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies. 19(3): 387-399.

Hassan L. A. H., Mahmassani H. S. & Chen Y. 2020. Reinforcement learning framework for freight demand forecasting to support operational planning decisions. Transportation Research Part E: Logistics and Transportation Review. 137, 101926.

Adenigbo J. A. 2016. Factors influencing cargo agent’s choice of operations in Abuja airport, Nigeria. Journal of Air Transport Management. 55, 113-119.

Chang Y. H. & Chang Y. W. 2009. Air cargo expansion and economic growth: Finding the empirical link. Journal of Air Transport Management. 15(5): 264-265.

Schafer J. L. & Graham J. W. 2002. Missing data: our view of the state of the art. Psychological methods. 7(2): 147.

Graham J. W. 2009. Missing data analysis: Making it work in the real world. Annual review of psychology. 60, 549-576.

Josse J. 2016. Contribution to missing values & principal component methods (Doctoral dissertation).

Rubin D. B. 1976. Inference and missing data. Biometrika. 63(3): 581-592.

Huisman, M., & Krause, R. W. (2018). Imputation of Missing Network Data.

Pedersen A. B., Mikkelsen E. M., Cronin-Fenton D., Kristensen N. R., Pham T. M., Pedersen L. & Petersen, I. 2017. Missing data and multiple imputation in clinical epidemiological research. Clinical epidemiology. 9, 157.

Dempster A. & Rubin D. 1983. Incomplete data in sample surveys. Sample surveys. 2, 3-10.

Enders C. K. 2017. Multiple imputation as a flexible tool for missing data handling in clinical research. Behaviour research and therapy. 98, 4-18.

Chen Q., Williams S. Z., Liu Y., Chihuri S. T. & Li G. 2018. Multiple imputation of missing marijuana data in the Fatality Analysis Reporting System using a Bayesian multilevel model. Accident Analysis & Prevention. 120, 262-269.

Audigier V., Husson F. & Josse J. 2016. Multiple imputation for continuous variables using a Bayesian principal component analysis. Journal of statistical computation and simulation. 86(11): 2140-2156.

Josse J., Pagès J. & Husson F. 2011. Multiple imputation in principal component analysis. Advances in data analysis and classification. 5(3): 231-246.

Josse J. & Husson F. 2012. Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique. 153(2): 79-99.

Tranmer M. & Elliot M. 2008. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 5, 30-35.

Harel O. 2009. The estimation of R 2 and adjusted R 2 in incomplete data sets using multiple imputation. Journal of Applied Statistics. 36(10): 1109-1118.