Time Series Analysis, Logistic Regression and Principal Component Analysis

Ananya Pratap Singh Chandel
School of Computing
National College of Ireland
Dublin, Ireland
x19237529@student.ncirl.ie

***Abstract*—The objective for this analysis is to apply three different time series models to a time series object and forecast up-to three periods ahead and compare their performance. The other objective is to apply and analyse binary logistic regression on data and evaluate the results. Lastly, the objective is to discuss Principal Component Analysis with the help of an example.**

***Index Terms*—Binary Logistic Regression, Time series analysis, Principal Component Analysis(PCA)**

I. Introduction

This Analysis focuses on applying the time series model on the energy consumption data for a country Estonia and trying to forecast the energy consumption for the next three years based on the time series data recorded annually over past 28 years i.e from 1990-2018. This could be useful in many ways like estimating the energy requirement for the coming years and being prepared for the future. Secondly, Binary logistic regression is performed on consumer data captured from a survey and its used to predict that weather a person will buy online music or not. This could be helpful for the companies selling online music to target the right customer base based on their attributes determined by applying Binary Logistic regression. Lastly, The concept of Dimensionality Reduction, Principal Component Analysis is discussed with the help of an example using US nutrition data in which a large set of variables are compressed to four factors that explain the underlying data without loosing much information. This could be helpful while applying Machine learning techniques.

II. Time Series Analysis

A. Introduction

The variables in cross sectional data are measured for single point in time while longitudinal data measures variables repeatedly over time. The time series, is series of observation for a variable over successive time period it can be taken every year, quarter, month, day, or any regular interval. Here time series analysis was performed on the energy consumption data of several countries recorded annually. The energy consumption data for the country Estonia is used which is recorded annually from the year 1990 to 2018. By applying time series analysis on the energy consumption for the country. The forecasts for the next three years is predicted using three different time series models and also comparing their performance.[4]

B. Objective and Source of data

Time series analysis is performed on data for energy consumption for Estonia from 1990-2018. Three different time series models are applied on the time series data and analysed to find the optimum model. The optimum model is used to forecast the energy consumption for three consecutive years.

The data is collected from the eurostat website the data is then filtered to obtain the annual energy consumption for Period 1990-2018, consumption is recorded in Gigawatt-hour for the country Estonia.[1]

C. Model application and evaluation

1) The data-set is read from an excel and loaded in R as a data-frame named energy and a time series object is created for further analysis. As seen in figure 1.

```
#read the excel file
energy <- read_excel("C:/Users/anany/Desktop/STATS CA2/arima.xlsx")
#convert the data into timeseries data object
tenergy <- ts(energy$Estonia, start=1990, frequency=1)
tenergy
```

Fig. 1. Time series object

2) The first step is plotting the time series data in order to check and analyze visually to check the data for Trend, Seasonality, Cyclical component or irregular fluctuations. A data is said to be stationary if it lacks all four. On plotting the time series object tenergy, it is concluded that the data is not stationary as a gradual trend can be seen in the data. As seen in figure 2.

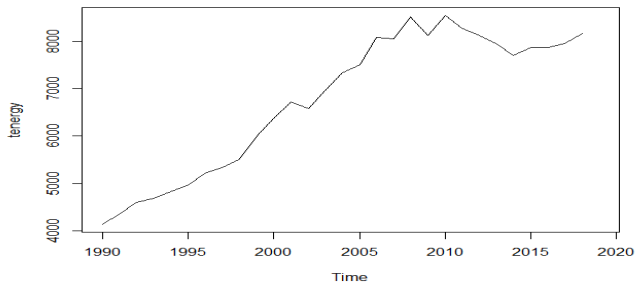

Fig. 2. Time series object Plot

3) Model1: Since, it could be seen from figure 2 that there is trend in the data Holt's model can be used for the time series analysis. A holt's model is an extended simple exponential model that allows us to forecast data with a trend. It creates a forecasting equation and two smoothing equations for trend and level.

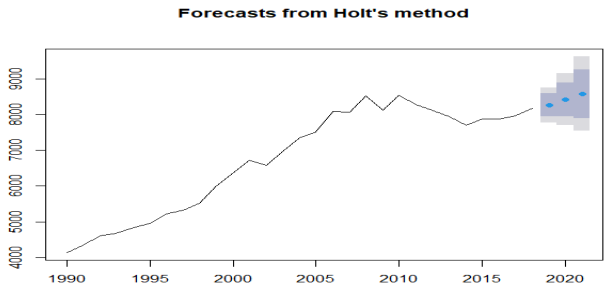

Fig. 3. holt's model forecast

The forecasts by the holt's model can be seen in figure 3. The resulting forecasts can be observed to have increasing trend and look sensible. The wide prediction intervals reflect the variation in the historical data. Further to analyse the fit of the model Ljung-Box test was performed, this test provides the test for autocorrelations to be zero. The results are insignificant as can be seen in figure 4. Hence, it can be stated that autocorrelations differ from zero and our model fits the data well.

On checking the accuracy for the model1. We see the RMSE to be 233.0577 and on plotting the residuals, A normal distribution is observed which provides evidence for a decent fit of the model as can be seen from figure 4 and 5.

```
> fit1
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2019       8264.899 7943.216 8586.582 7772.927 8756.871
2020       8422.219 7948.026 8896.411 7697.004 9147.433
2021       8579.538 7898.768 9260.309 7538.389 9620.687
> Box.test(fit1$residuals, type="Ljung-Box")

        Box-Ljung test

data:  fit1$residuals
X-squared = 0.2982, df = 1, p-value = 0.585

> #checkresiduals(fit1)
> accuracy(fit1)
                     ME     RMSE      MAE          MPE     MAPE      MASE        ACF1
Training set -1.74436 233.0577 173.3543 -0.003575759 2.430447 0.7060576 -0.09637203
> |
```

Fig. 4. holt's model accuracy

4) Model2: ARIMA models also known as Autoregressive integrated moving average models. While Exponential smoothing models are based on seasonality and trend. ARIMA models target to describe autocorrelations in data. ARIMA is used for forecasting time series data in which predictive values are the linear function of the actual values and residuals. The auto regressive(AR) terms contains the lag of stationary series in forecasting equation, While moving average(MA) terms consists of the lags in past forecast errors. If the time series is not stationary like in our data we see a trend the time series

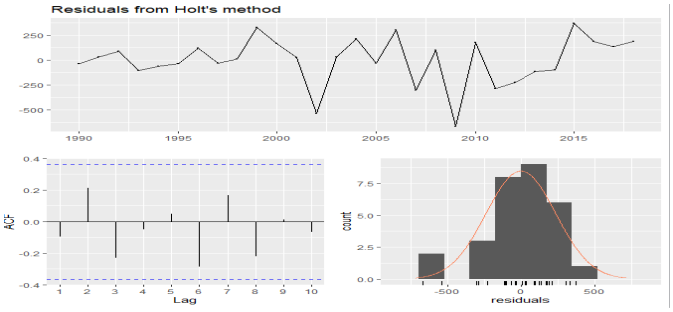

Fig. 5. holt's model residual

needs to be differenced in order to make it stationary, this constitutes the Integrated term(I) in ARIMA models. An auto ARIMA function in R provides the best fit for an ARIMA model. The Auto ARIMA function is applied on the data to obtain optimised value of the p, d, q terms where p is the auto-regressive element, d is an integrated element and q is the moving average element. On application of an auto ARIMA function, ARIMA(1,2,0) is suggested by the software. Ljung-Box test was performed on the model to check for the fit of model. The insignificant result of p vslur bring 0.26 shows the model fits the data well. On checking the plots for the residuals the plot seem to be a normal curve validating a decent fit for the model with an improved RMSE of 227.944 as can be seen in figure 6 and 7.

```
> fit2
Series: tenergy
ARIMA(1,2,0)

Coefficients:
          ar1
      -0.7454
s.e.   0.1169

sigma^2 estimated as 57954:  log likelihood=-186.27
AIC=376.53   AICc=377.03   BIC=379.13
> Box.test(fit2$residuals, type="Ljung-Box")

        Box-Ljung test

data:  fit2$residuals
X-squared = 1.2305, df = 1, p-value = 0.2673

> accuracy(fit2)
                     ME     RMSE      MAE          MPE     MAPE      MASE        ACF1
Training set -3.596851 227.944 171.8262 -0.05310999 2.417273 0.6998335 -0.1957653
> |
```

Fig. 6. Auto-ARIMA model accuracy

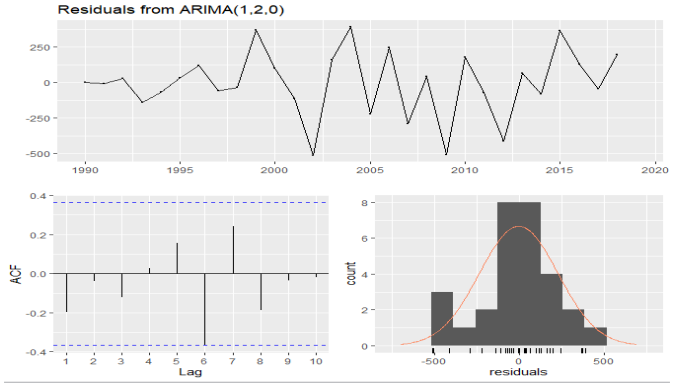

Fig. 7. ARIMA(1,2,0) models residual

5) Model3:Lastly, Different combinations of ARIMA models with different values of (p,d,q) are applied and the best model is selected as per the principal of parsimony. The first step while applying ARIMA is checking for stationarity i.e The properties of time series shall not depend on the time at which it is observed. As there is a trend in the time series ARIMA cant be applied. So to apply the time series model to the non stationary time series, We apply differencing. The Differencing involves the differences between the consecutive observations to be taken for making the time series stationary. Differencing is applied until a stationary time series is obtained. Here second order differencing worked. The level of differencing applied determines the Integrated element term and Hereby, the value of d for our ARIMA model is 2. The second order time series after applying differencing can be seen in figure 8. Augmented Dickey-Fuller test is applied to evaluate the stationarity of the time series. A significant test result of p value 0.01 suggests stationarity as can be seen from figure 9.

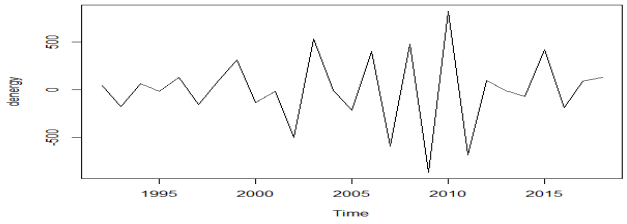

Fig. 8. After differencing

```
> adf.test(denergy)

        Augmented Dickey-Fuller Test

data:  denergy
Dickey-Fuller = -4.3093, Lag order = 2, p-value = 0.01216
alternative hypothesis: stationary
```

Fig. 9. ADF test

There can be autocorrelations in a time series which comprise of correlations with its own values from the past. An ACF plot also known as correlogram can be plotted to check the correlations. The correlations significantly away from 0 are represented by dotted line. As can be seen in figure 10.
Similarly, A Partial Autocorrelation at lag k can be plotted which shows the amount of correlation between x and x_lagk which is not explained by lower order autocorrelations. As can be seen in the figure 11. By looking at the ACF and PACF plots values for p and q are selected for the model. As d has already been set to 2 by looking at differencing. Similarly, looking at the ACF plot we see a significant autocorrelation at lag 1 and then

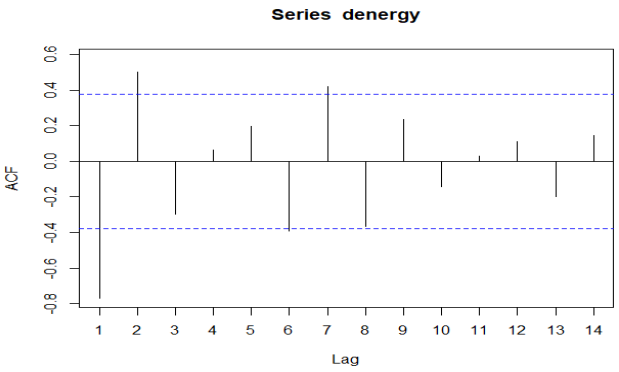

Fig. 10. ACF plot

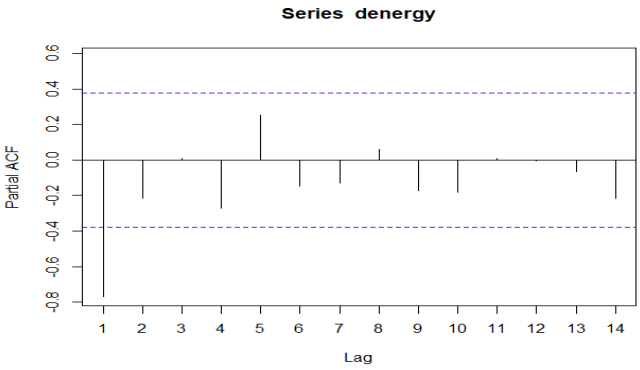

Fig. 11. PACF plot

the autocorrelations trail to 0 as lag increases. Similarly, looking at the PACF plots different patterns of ACF and PACF are identified by matching them with idealized patterns. Different values and combinations of p, d and q are tried based on the ACF and PACF plots until a model that fits the data most effectively is obtained.
An arima(p,d,q) model, With ARIMA(1,2,1) was fitted in with the data and an RMSE of 221.192 was obtained which appeared to be the lowest when compared to other two models. As can be seen in Figure 12. ARIMA(1,2,1) appears to be the most accurate model amongst the models are applied in this analysis. A Ljung box test was also performed to check the fit of the model that suggests that the autocorrelations do not differ from 0 as can be seen in figure 12. This ARIMA model appears to fit the data well as the residuals are seen to be normally distributed and the autocorrelations are zero for all lags and residuals have no relationship between them. It can also be seen from figure 13.

D. Evaluation and Forecast

Three different time series models have been applied for our time series analysis. The data is taken for the energy consumption for country Estonia for period 1990-2018. Model1 the holt's model was able to to forecast the data well with the accuracy and RMSE of 233.0577 While the auto ARIMA function

```
> #Fitting an ARIMA(p,d,q) model
> fit3 <- arima(tenergy, order=c(1,2,1))
> fit3

Call:
arima(x = tenergy, order = c(1, 2, 1))

Coefficients:
          ar1      ma1
      -0.6158  -0.3317
s.e.   0.1929   0.2700

sigma^2 estimated as 52549:  log likelihood = -185.53,  aic = 377.06
> Box.test(fit3$residuals, type="Ljung-Box")

        Box-Ljung test

data:  fit3$residuals
X-squared = 0.020306, df = 1, p-value = 0.8867

> accuracy(fit3)
                     ME    RMSE      MAE        MPE     MAPE     MASE         ACF1
Training set -7.870746 221.192 161.5376 -0.1061446 2.268573 0.657929 -0.02514863
```

Fig. 12. Arima(1,2,1) Accuracy

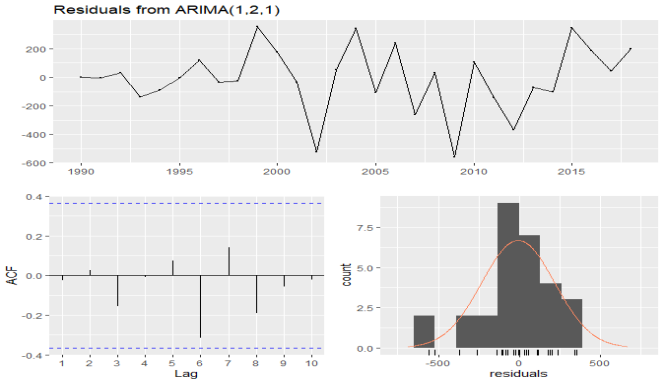

Fig. 13. Arima(1,2,1) model Fit

suggested fitting ARIMA(1,2,0) with the accuracy and RMSE of 227.944. Lastly, Different variations of ARIMA(p,d,q) model were applied, based on the different values of p, d and q. It was Concluded that keeping the principal of parsimony in mind the arima model ARIMA(1,2,1) fits the data best with the improved accuracy and RMSE of 221.192. Among the other three fitted models ARIMA(1,2,1) gave the least RMSE value and showed normal distribution of residuals, Hence its selected to fit the data well when compared to the other two models.

Therefore ARIMA(1,2,1) model is used to forecast the energy consumption for Estonia for three periods ahead that is for the year 2019, 2020 and 2021. As can be seen in figure 14 and 15. The resulting forecasts can be observed to have increasing trend and look sensible. The wide prediction intervals reflect the variation in the historical data.

```
> #Forecasting with the fitted ARIMA(1,2,1)model
> forecast(fit3, 3)
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2019       8242.827 7949.050 8536.604 7793.534 8692.121
2020       8401.606 7975.097 8828.114 7749.317 9053.894
2021       8505.129 7850.941 9159.317 7504.634 9505.624
> plot(forecast(fit3, 3), xlab="Year", ylab="Energy Consumption")
```

Fig. 14. Arima(1,2,1) Forecast

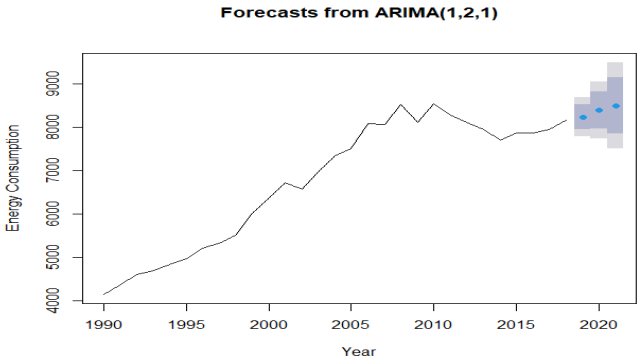

Fig. 15. Arima(1,2,1) Forecast Plot for 3 periods

E. Conclusion

It can be concluded after applying three models, model1(Holt's model), Model2(auto-ARIMA(1,2,0)), Model3(ARIMA(1,2,1)). The Model3 best fits the data for energy consumption of Estonia with the least RMSE value of 221.192. And is therefore selected to forecast the energy consumption for the next 3periods(years) for Estonia. The forecast values and prediction intervals can be seen in figure 14 and the graph or plot can be seen in figure 15. This forecast-ed data for the year 2020, 2021 and 2022 can be useful for the Energy generating companies in Estonia as it can help them determine the future requirement and prepare to fulfill it well in advance.

III. Binary Logistic Regression Analysis

A. Introduction

A logistic regression is applied when the dependent variable is dichotomous or is having two classes, it can use both quantitative and qualitative variables as independent variables. Like other regression analysis Binary logistic regression is also a predictive analysis. It can predict the likelihood of an event to fall into either of the two classes of dependent variable.

B. Objective

The objective for this analysis is predicting whether, A Consumer will buy an online music subscription based of certain parameters that are taken from a survey for Consumer Choice.

C. Data Description

The data used for the analysis is taken from a Consumer Survey taken from Pew research Centre website.[2] The data-set was cleaned and unwanted columns were removed and loaded in SPSS. A sample size of N=195 is taken. The sample contains the following variables.

Dependent variable:-

boughtmusic: its a dichotomous variable with 'yes' for people who bought music online 'no' for who didnt.

Independent Variable:-

Age: its a quantitative variable that contains the age of the people taking part in the survey.

Sex: its a nominal variable with Male/Female values

useinternet: its a nominal variable with Yes/No values for people who use internet or not

internetbanking: its a nominal variable with Yes/No values to state that whether the person uses internet banking or not.

D. Software used for Analysis

SPSS by IBM, 26th version

E. Assumptions for Analysis

1) Sample-size:
Binary Logistic regression works more efficiently when the sample sizes are sufficiently large, there should be atleast 20 cases per predictor for our analysis(min 80 cases in total). The data used in this analysis has 195 rows and 4predictors. It can be seen in figure 16. This meets the criteria of having an adequate sample size hence meeting the assumption.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	195	100.0
	Missing Cases	0	.0
	Total	195	100.0
Unselected Cases		0	.0
Total		195	100.0

a. If weight is in effect, see classification table for the total number of cases.

Fig. 16. Sample size

2) Dependent Variable outcome is mutually exclusive:
From figure 16 and 17 it can be seen that dependent variable has two classes which contains values as Yes and No and account to be N i.e 195.

Classification Table[a,b]

			Predicted		
			bought music		Percentage Correct
	Observed		0	1	
Step 0	bought music	0	0	72	.0
		1	0	123	100.0
	Overall Percentage				63.1

a. Constant is included in the model.

b. The cut value is .500

Fig. 17. block o classification table

F. Evaluation

1) Block 0:
The block 0 represents the null model without independent variables. From figure 2.2 it can be seen that 63% of the people buy music online.
2) Block 1, Omnibus test:
This test holds the null hypothesis, H0 to be stating that the coefficients of the predictors is 0. For this Analysis the sig. or p value less than 0.05 and Chi-square value of 29.671 with degrees of freedom as 3 is obtained, As can be seen in figure 18. Thereby H0 is rejected. This suggests our model and independent variables are suitable to predict that someone will buy online music or not.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	29.671	3	.000
	Block	29.671	3	.000
	Model	29.671	3	.000

Fig. 18. Omnibus test

3) Model Summary:
From figure 19 it can be seen that the value for Cox and Snell R square is 0.141 for this model and the value for Nagelkerke R$\hat{2}$ is 0.193 this tells us that 14% to 19% of variability in target variable can be explained by the model.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	227.161[a]	.141	.193

Fig. 19. Model Summary

4) Hosmer and Lemeshow test:
This test is performed to test the fit of model. If the value of sig less than 0.05 this indicated that the fit is poor. The significance of 0.821 shows the fit is not poor and indicates support for model. This can be seen in figure 20.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.384	8	.821

Fig. 20. Hosmer and Lemeshow test:

5) Classification Table:
From the figure 21 it can be seen that the percentage accuracy in classification is 68.2% It can be therefore inferred that the model correctly predicted or classified 68.2% of people who will buy music online.
6) Variable in The Equation:
   a) Referring to figure 22 we check the Wald statistics its similar to t-statistic in regression for Wald statistic the null hypothesis states that B=0. Hence the p value tells us that whether our predictors are significant or not the value of p less than 0.05 tells

Classification Table[a]

Observed			Predicted: bought music 0	Predicted: bought music 1	Percentage Correct
Step 1	bought music	0	27	45	37.5
		1	17	106	86.2
	Overall Percentage				68.2

a. The cut value is .500

Fig. 21. Classification Table

us that the predictors are significant for our model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	use internet	1.296	.579	5.002	1	.025	3.653
	internetbanking	.771	.326	5.608	1	.018	2.162
	AGE	-.029	.010	8.865	1	.003	.971
	Constant	.345	.773	.200	1	.655	1.413

a. Variable(s) entered on step 1: use internet, internetbanking, AGE.

Fig. 22. Variables in the Equation

b) The predictor Sex was removed as it resulted in a non significant value for our model hence we only use three independent variables that are significant(p less than 0.05) to predict our dependent variable i.e Age, use of internet and whether a person uses internet banking is used to predict our dependent variable that whether if a person buys music online.

c) The equation for logistic regression can be summarised as follows:
Log(p/1-p)=0.345+1.296(useinternet)+0.771 (internet banking)-0.029(AGE)

d) The odds ratios are depicted as Exp(B). As can be seen from figure 22. The use of internet shows the highest odds ratio of 3.653 which tells us that the odds of person buying music online increases by 3.653 times if he uses internet. Age shows negative value showing an inverse relationship with the dependent variable.

G. Conclusion

By looking at the results obtained from this Binary logistic regression model, it is concluded that our independent variables Age, Use and access to internet and internet banking can predict or determine how likely he is to buy music online which is our dependent variable with an accuracy of 68.2%. This information can be used by the companies selling music online to target a specific consumer base and help increase their sales.

IV. Principal Component Analysis

A. Introduction

Principal component analysis(PCA) is a method for dimensionality reduction. It is used for reducing the dimensionality of large datasets by trimming down large set of variables to a small set of variables. These reduced set of variables are transformed in such a way that they contain most of the information contained by the large set of variables at the same time being uncorrelated to each other. Transforming large set of variables into small set of variables increases simplicity for our analysis but this comes at the expense of accuracy. Dimensionality reduction involves a tradeoff of accuracy for simplicity. While working on large datasets for machine learning to analyze and visualize the data, dimensionality reduction can help in making the process faster by significantly reducing the extraneous variables to process without loosing much information. PCA constructs principal components by transforming the correlated variables without much loss of important information. The principal components show maximum variance and are uncorrelated.

B. PCA example

1) The example is taken from USDA National Nutrient Database it contains standard reference for food composition data for US. The data was cleaned to obtain 35columns and 1500rows. SPSS is used as a software to perform principal component analysis. Sample size should be adequate to perform PCA at least 10-20 observations per variable. Hence this data meets the criteria.[3]

2) The relationship of the variables is checked carefully there should be adequate correlations among the variables for the factor analysis to provide useful information. There is adequate correlation among the variables in the example as can be seen by figure 23. The correlation matrix shows that variables have high correlation and PCA can be applied.

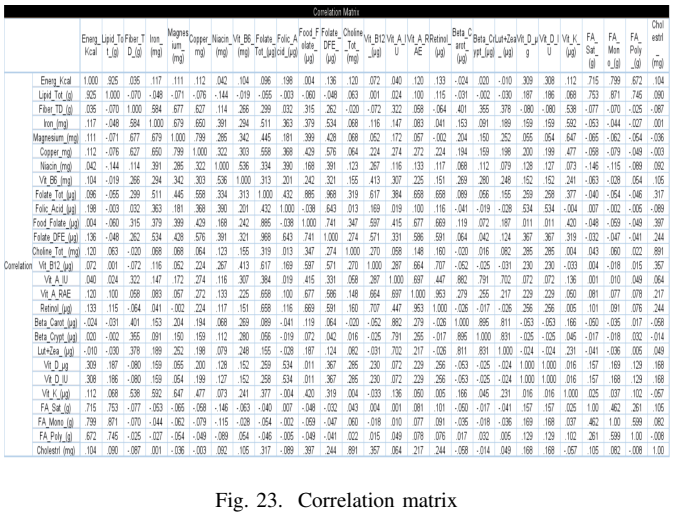

Fig. 23. Correlation matrix

3) Correlations coefficients are inspected to be greater than 0.3 and Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin (KMO) criteria are checked for the data. Bartlett's Test of Sphericity is a null hypothesis test which states that correlations between the data set variables are 0. For the example data to be applicable the null hypothesis has to be rejected. A significant p-value rejects the null hypothesis and states that there is a correlation pattern among the variables and PCA can be performed. The Kaiser-Meyer-Olkin is a measure of sampling adequacy, its maximum value is one for a good factor analysis KMO should exceed 0.5. The example data fulfills both the criteria as can be seen in figure 24.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.695
Bartlett's Test of Sphericity	Approx. Chi-Square	66004.257
	df	351
	Sig.	.000

Fig. 24. Bartlett's Test of Sphericity and KMO

4) The data needs to be standardized before performing PCA. As the analysis is performed in SPSS the correlation matrix function first standardizes the variables and then factors the co variance matrix so for this example we need not standardize the data.
5) The factors obtained from PCA shall explain maximum variance in the data by using as few components as it can. Choosing factors is based on Kaiser's Criterion which states that the factors with eigenvalue greater than 1.0 or more shall be taken into consideration, Eigen values represent the amount by which the factor can explain the variance . Catell's scree test can also assist in choosing factors/components by looking at the scree plot usually the factors above the elbow break are retained. For our example the total variance can be shown by 4 components as seen from the figure 26. The four components with Eigenvalue greater than 2.909 explain 64.913% of variance. By looking at the scree plot from figure 26 it can be cross-validated and verified that there are only 4 factors above the elbow point after which there is not much difference among the other factors Eigen value.
6) As it can be seen from the figure 25 There are 7 components with Eigen value greater than 1 and they can be selected as principal components. But we only choose 4 components from our Analysis as the components with Eigen values less than 2.909 showed very less variance. For our example from the figure 25 we see that there are significantly 4 components above the break elbow rest of the points in the scree plots show very less variance. The 4 components selected alone explain 64.913% of variance for our example.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.954	25.757	25.757	6.954	25.757	25.757	4.454	16.497	16.497
2	4.217	15.618	41.375	4.217	15.618	41.375	4.122	15.266	31.763
3	3.446	12.764	54.139	3.446	12.764	54.139	3.853	14.271	46.033
4	2.909	10.775	64.913	2.909	10.775	64.913	3.618	13.402	59.435
5	2.087	7.729	72.642	2.087	7.729	72.642	2.682	9.934	69.369
6	1.688	6.250	78.892	1.688	6.250	78.892	2.050	7.594	76.963
7	1.237	4.582	83.474	1.237	4.582	83.474	1.758	6.511	83.474
8	.807	2.990	86.465						
9	.723	2.678	89.142						
10	.558	2.066	91.208						
11	.440	1.630	92.838						
12	.347	1.285	94.123						
13	.335	1.241	95.364						
14	.300	1.109	96.473						
15	.246	.910	97.384						
16	.183	.678	98.062						
17	.174	.643	98.705						
18	.124	.458	99.163						
19	.096	.357	99.520						
20	.078	.289	99.809						
21	.049	.183	99.992						
22	.002	.006	99.997						
23	.001	.002	100.000						
24	6.450E-5	.000	100.000						
25	8.026E-6	2.973E-5	100.000						
26	4.868E-6	1.803E-5	100.000						
27	5.992E-7	2.219E-6	100.000						

Extraction Method: Principal Component Analysis.

Fig. 25. Total variance and Eigen value

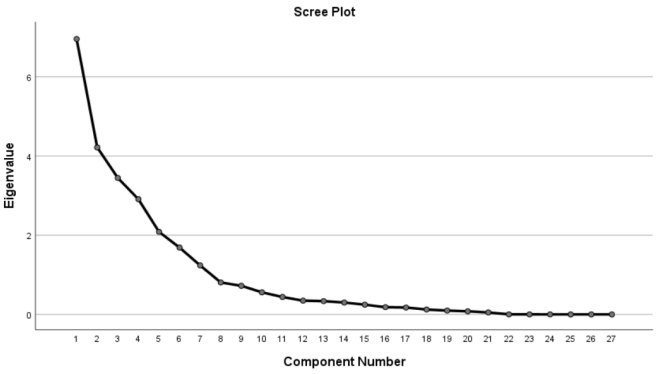

Fig. 26. Scree Plot

7) For making the components more interpret-able a rotation technique is performed. Rotation can be used to enhance the correlations between the factor/components. In this example we used Varimax rotation as can be seen in figure 27. Varimax is an orthogonal rotation which helps to purify the loading matrix columns.

Rotated Component Matrix[a]

	Component						
	1	2	3	4	5	6	7
Retinol_(µg)	.949						
Vit_A_RAE	.915						
Food_Folate_(µg)	.791						
Vit_B12_(µg)	.782						
Folate_Tot_(µg)	.767						
Folate_DFE_(µg)	.682						
Magnesium_(mg)		.886					
Iron_(mg)		.806					
Vit_K_(µg)		.805					
Fiber_TD_(g)		.783					
Copper_(mg)		.775					
Lipid_Tot_(g)			.993				
Energ_Kcal			.939				
FA_Mono_(g)			.872				
FA_Poly_(g)			.770				
FA_Sat_(g)			.718				
Beta_Carot_(µg)				.957			
Beta_Crypt_(µg)				.949			
Lut+Zea_ (µg)				.875			
Vit_A_IU				.861			
Vit_D_IU					.923		
Vit_D_µg					.923		
Folic_Acid_(µg)					.743		
Choline_Tot_ (mg)						.944	
Cholestrl_(mg)						.936	
Niacin_(mg)							.844
Vit_B6_(mg)							.759

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Fig. 27. Varimax rotated components

8) Before the PCA was performed the example contained 27 variables with more or less correlation. After the PCA is successfully performed on the data the variables are reduced to 4 factors that explain 64.913% of the

variance.

9) After performing Principal component analysis on data with 27 variables we are down to just four factors that explain the maximum variability. These components can be named as:
   Factor 1 can be called as Food composition rich in Vitamins(Vitamin A, Vitamin B12 and Folate).
   Factor 2 can be called as Food composition rich in Minerals(Magnesium iron fibre copper).
   Factor 3 can be called as Food composition rich in Energy and Fats.
   Factor 4 can be called as Food compositions rich in beta Carot.

V. REFERENCES

[1]"Supply, transformation and consumption of electricity",Ec.europa.eu, 2020. . Available:https://ec.europa.eu/eurostat/databrowser/view/nrg_cb_e/default/table?lang=en.

[2]"September 2007 – Consumer Choice", Pew Research Center: Internet, Science Tech, 2021. [Online]. Available: https://www.pewresearch.org/internet/dataset/september-2007-consumer-choice/.

[3]"USDA National Nutrient Database for Standard Reference", 2021. [Online]. Available: https://data.nal.usda.gov/dataset/usda-national-nutrient-database-standard-reference-legacy-release

[4]R. Hyndman and G. Athanasopoulos, Forecasting. 2020.