Principal Component Analysis on Average Retail Price of Electricity – Case Study

Arman Sadreddin, Hadi Shahidi Nejad, Andrea Schiffauerova

Concordia Institute of Information Systems Engineering, Concordia University, Montreal, QC, Canada

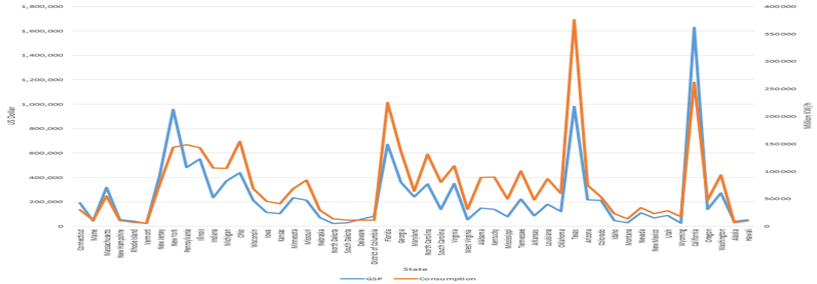
Introduction

Abstract

This research contains measurements of average retail price of electricity to ultimate customers by End-User Sectors, Residential, Commercial, Industrial and Transportation sectors, in 52 states of United States of America. Principal component analysis procedure was used to analyze the correlation between the amounts of electricity price of each sector for each state based on the 2005 USA states' electricity consumption data-set. This result can help the decision makers to handle the future issues related to price and consumption in power industry.

Energy Information administration yearly collect the reading of average price of electricity been charged to different sectors in different states. It helps to make the annual electric sales and revenue generated reports. Based on the analysis of the data, decision maker decides the price of electricity to be charged to different sectors in different states to increase their revenue and control the consumption.

Figure below shows the amount of consumption of the electricity in each state of USA and GSP (Gross State Product).



Objectives

- ✓ The examination of the correlations between the variables of a selected set
- ✓ The reduction of the basic dimensions of the variability in the measured set to the smallest number of meaningful dimension
- ✓ The elimination of variables which contribute relatively little extra information
- ✓ The examination of the grouping of individuals in n-dimensional space

Methodology

Principal Component Analysis

Principal components analysis (PCA) is an explanatory technique to learn about data sets. The objective of PCA is to reduce the dimensionality of the data set while retaining as much as possible the variation in the data set.

Principal components (PCs) are linear transformations of the original set of variables, and are uncorrelated and ordered so that the first few components carry most of the variation in the original data set.

Given a data matrix X, the PCA algorithm consists of four main steps:

PCA Step#1: Calculate adjusted dataset X of size n*p

PCA Step#2: Calculate co variance matrix S

PCA Step#3: Calculate eigenvectors/eigenvalues of S

PCA Step#4: Calculate transforming dataset Z=XA, where A is the matrix containing the eigenvectors.

Results (Contd.)

In this project our dataset contains measurements of electricity consumption for Residential, Commercial, Industrial and Transportation sectors as this columns:

R2005: Residential Sector

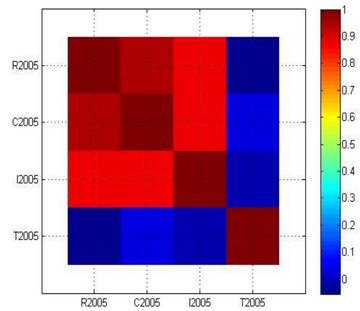
C2005: Commercial Sector

I2005: Industrial Sector

T2005: Transportation Sector

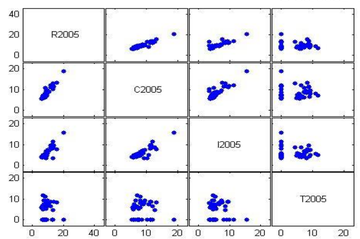
Correlation Matrix , Correlation Matrix based on Color

	R2005	C2005	I2005	T2005
R2005	1.000	0.947	0.869	-0.055
C2005	0.947	1.000	0.872	0.039
I2005	0.869	0.872	1.000	-0.013
T2005	-0.055	0.039	-0.013	1.000



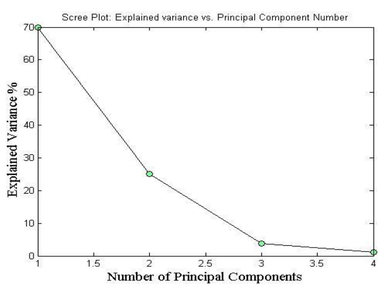
Average Retail Price of Electricity to Ultimate Customers in Residential sector is highly correlated with the Commercial sector, and then it's correlated to the Industrial sector and at last the correlation with transportation sector is negative by -0.104.

Scatter Plot Matrix

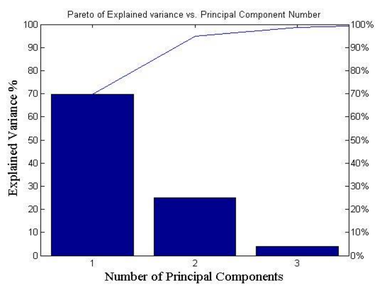


A scatter plot matrix shows relationships among several variables taken two at a time. Scatter plot matrices can show a value of information, including direction, strength, and outliers.

Scree Plot



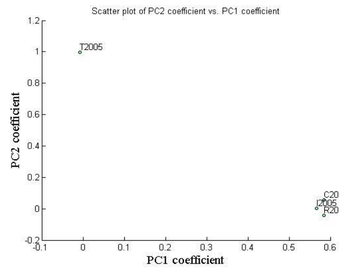
Pareto Chart



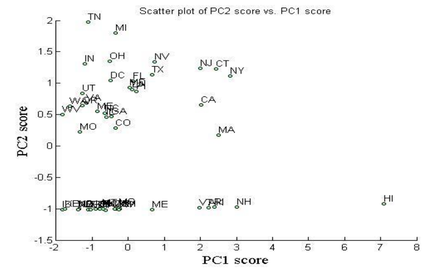
In above scree plot first two components are responsible for 80 % of variance. First components accounts to 70 % and second component accounts to 20 %.

Results (Contd.)

PC1 vs. PC2 Coefficient Plot



PC1 vs. PC2 Score Plot



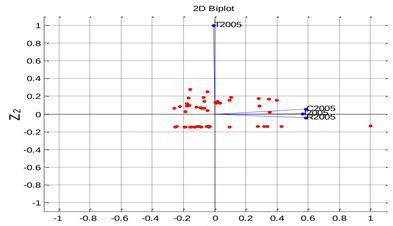
Eigenvector Matrix

0.97	-0.04	-0.16	-0.15
0.97	0.06	-0.15	0.16
0.95	0.00	0.32	-0.00
-0.02	1.00	0.00	-0.02

X1 = R2005
X2 = C2005
X3 = I2005
X4 = T2005

First PC is represented as:
Z1=0.97X1+0.97x2+0.95X3-0.02x4
Second PC is represented as:
Z2=-0.04x1+0.06x2+0.00x3+1.00X4

BiPlot – 2D



The Axis in the biplot shows the principal components (columns of eigenvector matrix), and the observed variables (rows of matrix) are represented as vectors. Each observation is represented in the biplot as red points and the location of these points shows the score of each observation for two principal components. Biplot helps to see the sign of each variable against both the components.

Conclusion & References

Conclusion:

It was difficult to analyses the multivariate data without the help of PCA. We were able to reduce the dimensionality of the data and we figured out which sector has high average price of electricity and where consumption of electricity was more. As per plots we made, we found that Residential, commercial and industrial electricity price was more in Hawaii and Michigan and Tennessee has highest price in transportation sector although residential and commercial were the main sectors whose price was high so we can ignore transportation sector. This data can help the decision maker to handle the future issues related to price and consumption in each state.

References:

[1] A. Ben Hamza, Lecture Notes: STATISTICAL PROCESS AND QUALITY CONTROL. Concordia University, Montreal Canada 2012

[2] A. Ben Hamza, Chapter 5: Statistical Process and Quality Control, Concordia University, Fall 2013

[3] Mathforum.org

[4] Mathworks.com