Applying Data Mining Techniques to Forecast the Natural Gas Consumption Using an Effective Business Intelligence Model.

Ben Bella S. Tawfik, Abdel-Fatah A. Hegazy, and Amro A. Shoeib

Arab Academy for Science and Technology and Maritime Transportation
Misr El-Gadida, Cairo, Egypt
Benbellat@gmail.com, Hegazy@aast.edu, Amro4cs@hotmail.com

Abstract

The key issue for decision making in public utility- Enterprises is obtaining the right information at the right time to the right decision makers. Here we present a model intended to predict mainly the residential, commercial and industrial natural gas consumption by applying BI/DM - a predictive analytics method on a certain predefined attributes using regression algorithm and a multiple nonlinear regression model which mathematically represents the relationship between natural gas consumption and influential variables. Mining the natural gas public utility-Enterprise Data Warehouse of an administrative district to assist in management planning for effective decision making. Data Preprocessing includes cluster analysis. Wavelet transform is used to analyze the data with different scale (multi-resolution analysis). In this phase, the selection of the best scale which describe the actual data and ignore the details (or noise). In this work different fitting models are introduced. The first model was polynomial fitting with single independent variable. Then multiple linear regression models are introduced. Finally, non linear regression models are discussed. To measure the goodness of fit, R2 value is calculated. The challenge was to select the best fitting model. The best model is reached with 4th order non-linear regression model with 0.97 R2 value.

Keywords: *BI, DSS, DM, Cluster Analysis, Wavelet Transform, Polynomial fitting, Multiple Regressions, Predictive Analytics.*

1. Introduction

The prediction of natural gas consumption is crucial for the supply and demand of the governmental and private agencies associated to the natural gas sector. In particular, the long range prediction, 1 to 5 years, is important to ensure the supply of natural gas to a given city or community. This type of prediction is particularly important for countries like Egypt, where the production sites are far from the major centers of consumption to ensure continuous supply of natural gas demand. This circumstance and the lack of large reservoirs make it necessary to develop reliable models to predict the gas consumption a few years in advance.

The present study addresses solely the issue of residential, commercial and industrial Natural Gas consumers due to their consumption variation patterns versus their location. The market of NG in Egypt is influenced by season demand. NGC in Egypt fluctuates from highs in months like Ramadan to lows during the summer vacations. These consumptions fluctuate in response to factors of which present Egyptians feasting habits, but do not appear to be the only factor. (These factors will not be taken into consideration).

There is also a need to predict the NGC in the intermediate range of time, within 1 to 2 years, in order to adapt and upgrade the infrastructure of transportation and distribution. This type of prediction is also useful for all the sectors of the gas industry that need to plan their production and optimize their anticipated purchase. The most important factors that affect the gas consumption of residential and commercial users are temperature, day-of-the-week (holiday or working day) and prevailing scenario of consumption.

Other factors that may also influence the consumption are: wind speed and its direction, humidity, etc. Due to the lack of reliable information on these parameters, we have not included them in our model. If these consumptions could be predicted with greater precision, efficiency gains could be realized.

There are several approaches to forecast the demand of natural gas based on different methods. Traditionally they were based on the concept of season's demands, multiple linear regressions, and more recently they were based on Decision trees, regression, and cluster analysis to form the core algorithms for most data miners. This has been very consistent over time [1].

We herein present the basic characteristics of simple models to perform long range prediction and its generalization for future predictions.

2. Model of Prediction

The goal of data mining is to extract knowledge from a data set in a human-understandable structure and considered the analysis step of the Knowledge Discovery in Databases process (KDD), a relatively young and interdisciplinary field of computer science [2].

Since Data Modeling (DM) is the process of discovering new patterns from large data sets involving statistics, database systems, data management, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structure, visualization and online updating. DM involves six common classes of tasks (Methods):

- **Anomaly detection**

  This step involves identifying the unusual data records that might be interesting or data errors and requiring further investigation (Outlier/change/deviation detection).

- **Association rule learning**

  This task/method searches for relationships between variables (Dependency modeling).

- **Clustering**

  This process is discovering groups and structures in the data those are in some way or another "similar", without using known structures in the data.

- **Classification**

  Here, the main objective is to generalize known structure to apply to new data.

- **Regression**

  This step includes many processes, first polynomial fitting, second, linear multiple regression, and finally non linear multiple regression analysis. The goodness of fit is measured using R2 value.

- **Summarization**

Finally provide a more compact representation of the data set, including visualization and report generation.

The outcome of statistical inference may be an answer to the question "what should be done next?", where this might be a decision about making further experiments or surveys, or about drawing a conclusion before implementing some new organizational or governmental policy.[3]

3. System Structure

3.1 Data collection and Capturing

Data is collected from different sources nationally in a structured, systematic and scientific way to provide information about the petroleum sector specifically the natural gas sector in Egypt from the ministry of petroleum and metallurgical wealth, Natural Gas Key Indicators observing years (2000/2001-2009/2010) , Central Agency for Public Mobilization And Statistics, The Ministry of health and population, Ministry of investments, Ahram center for petroleum and energy studies, Information and decision support center, and The National Authority for Remote Sensing And Space Sciences.

A historical data base from a public utility enterprise serving governorate of Sharkia was acquired, it held all consumers information about natural gas consumption over the period year 2002 through 2010.

This limited number of seasons was chosen for the fact that the data was more readily available.

3.2 Data Preprocessing

Is an often neglected but important step in the data mining process, the representation and quality of data is first and foremost before running an analysis. [4]

This includes cleaning, normalization, transformation, and multi-resolution data analysis. The product of data pre-processing is the final training set. [5]

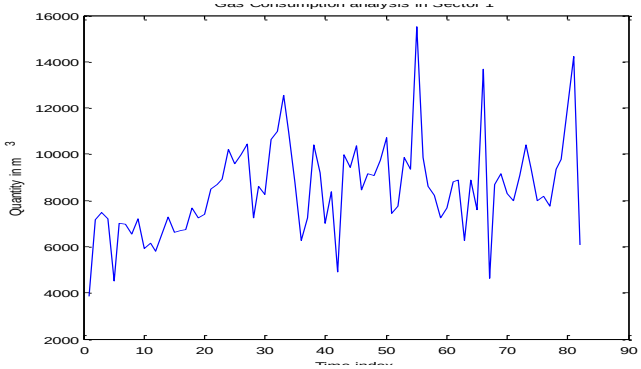
3.3 Extract, Transform and Load:

Before DM algorithms can be used, a target data set is assembled by Cleaning up any problems with the source data so that they are consistent and well organized by removing the observations with noise and missing data. Using MS SQL 2008 server & MATLAB as an analytical tool, we perform extraction and staging on both repository files <.MDF , .LTF>, export data to Excel-worksheet, and Convert to another format to ensure accuracy and be able to decompose NGC over repetitive months in different years and sectors then we load data to new data warehouse in a data structured array using MATLAB showed in Fig. 1, fig.2, and Fig. 3.

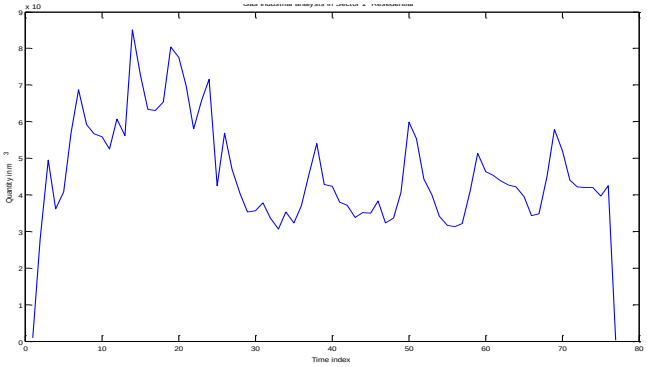
3.4 Cluster Analysis/Clustering:

It is a main task of explorative data mining, and a common technique for statistical data analysis and it is used in many fields, using Cluster Centroid model (hard/strict partitioning) clustering to define Market groups (Residential, Commercial, Industrial), Time period groups

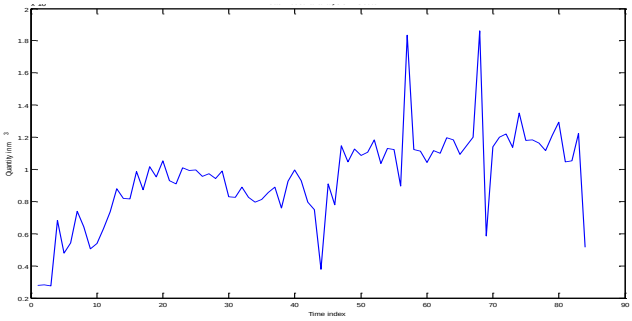
(Years, Months) and Sector groups (Sector1, Sector2,…,Sector6). Here we mean segmentation, Space is defined as default n-dimensional space, or is defined as above, or is a predefined space driven by past experience (unsupervised learning) accumulative historical database.



Figer 1: Data analysis on Residential Consumers in sector1.



Figer 2: Data analysis on Commersial Consumers in sector1.



Figer 3: Data analysis on Industrial Consumers in sector1.

3.5 Modeling

With the de-normalized records, we construct a new data warehouse with multi-dimensionality to facilitate the analysis of NGCs' behavior, by using a multi-dimensional model (Market D, Time D, and Location D). MS-SQL server 2008 as a tool to process the cube possessed from the DW repository, and taking into consideration the complex ERD data, there are <524159> rows of valid samples. [6]

Applying (BI-Algorithm(s) 1/2/3): To Test data set and Validating results to recorded validation data set.

3.5.1 Wavelet transform:

By Converting a signal/data into a series of wavelets to provide a way for analyzing waveforms, bounded in both frequency and duration allowing signals/data to be stored more efficiently and be able to better approximate real-world signals/data. The CWT,FT, and FFT provides an ideal apportunity to examine the NGC variaion occuring (Where and When) respectively (Sector and Time dimensions). [7] [8]

We construct a 2-dimensional picture of wavelet power showing the natural gas consumption quantity peaks in the spectrum $x(t)$ and how those peaks change with time in each sector (1,…, 6). [9] [10]

$$\mathrm{CWT}_x^{\psi}(\tau,s)=\Psi_x^{\psi}(\tau,s)=\frac{1}{\sqrt{|s|}}\int x(t)\bullet\psi^*\left(\frac{t-\tau}{s}\right)dt \quad (1)$$

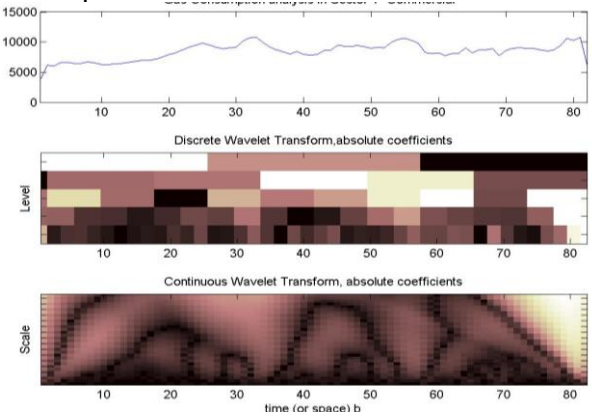
Where τ is the translation, (the location of window),

s is the scale, and

$\psi^*\left(\frac{t-\tau}{s}\right)dt$ is the mother wavelet.

After decomposing the test data set with wavelets, it is observed that data is non-stationary Signal/data, meaning that Frequency content changes in time. A basis of two months measurement of NGC is taken according to each sector; we index our measurement series to contain 82 data records i.e. the series scale range of 1:82 indices. [11][12]

Wavelet transform emphasizes the data description for different scales (frequency). It is shown in Fig. 4 the gas consumption variation with time for different scales. [13]



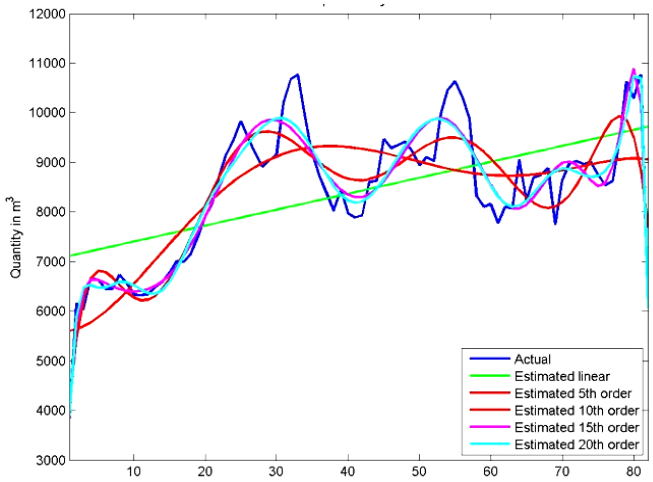
Figer 4: Wavelet Transform applied on Commercial Consumers in sector1

3.5.2 Polynomial Fitting

Fitted curves are used as an aid for data visualization. We construc a curve/mathematical function, that has the best fit to the series of data points previously analyzed Fig. 5, for each model (1st order 5th order 10th order, 15th order, and 20th order) R2 value to measure the

goodness of fit. The results are 0.3112, 0.6162, 0.7806, 0.8848, and 0.8934 in sequence. The higher R2, the better fitting.

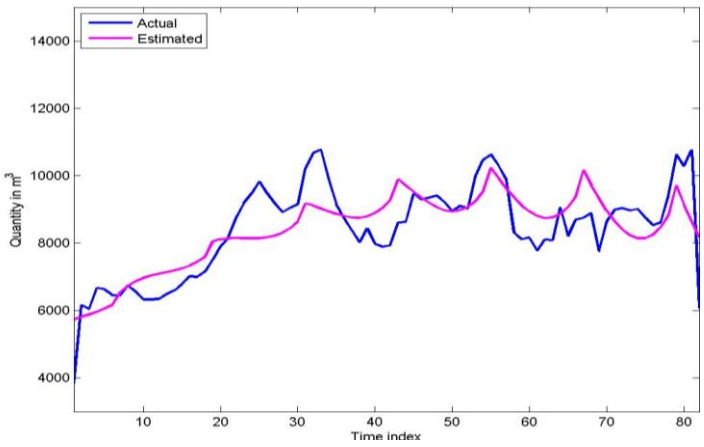
$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + a_nx^n \quad (3)$$



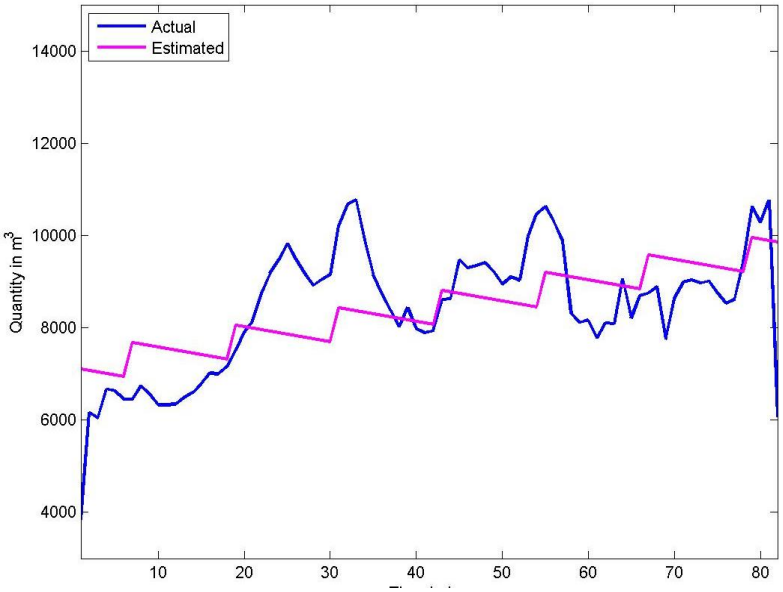
Figer 5: Polynomial Fitting applied on Residential Consumers in sector1.

3.5.3 Multiple Regression:

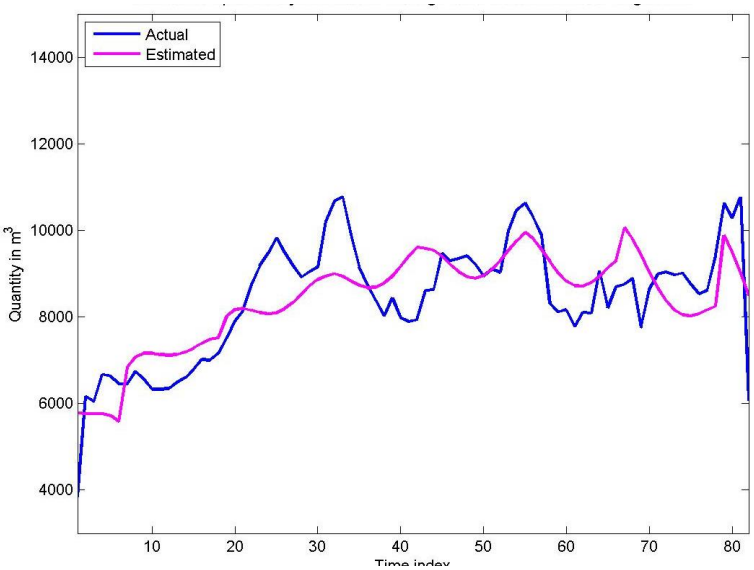
In this case, five cases are considered. First, multiple linear regression with two independent variables X1, and X2 for years and months in sequence. The main objective is to provide a tool for finding a solution for unknown coefficients β that will, minimize the distance between the measured and predicted values of the dependent variable Y also known as method of least squares. [14] [17] The model function has the form Y=f(X1, X2). The main objective is adjusting the parameters of a model function to best fit a data set applied on Residential Consumers in sector1 as a sample; results are shown in (Figure 6. Through Figure 10.). [15][16]



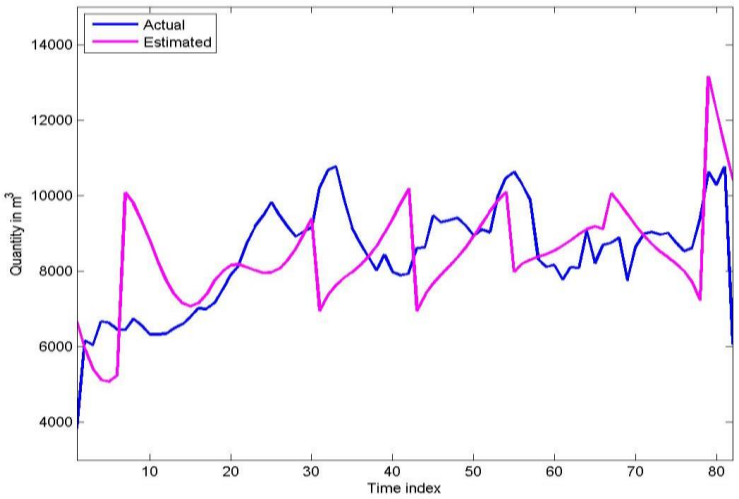
Figer 6: The 1st Order Non Linear Regression Model



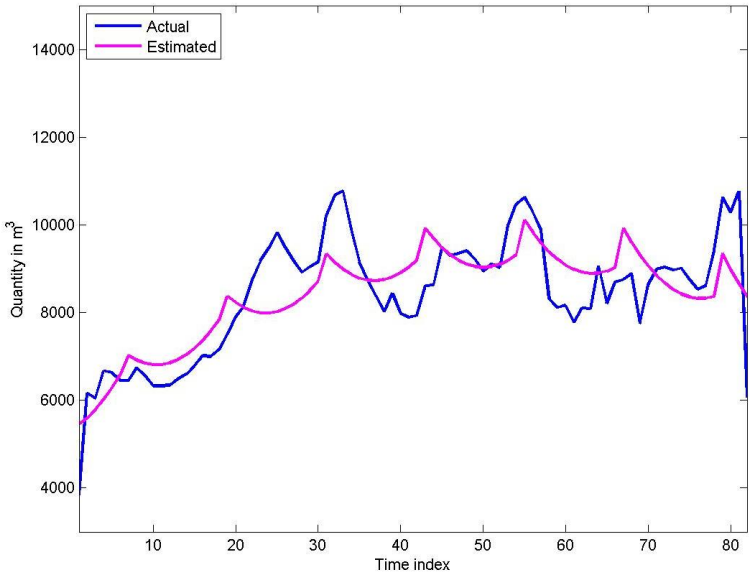
Figer 7: The 2nd Order Non Linear Regression Model



Figer 8: The 3rd Order Non Linear Regression Model



Figer 9: The 4th Order Non Linear Regression Model (First Model – 12 coefficients)



Figer 10: The 4th Order Non Linear Regression Model (Second Model – 14 coefficients)

From these results it is clear that the best fitting is reached (Figure 10.) with 4th order with only 14 coefficients using Non Linear Regression Model. The measure of goodness for the previous regression models are (R2) 0.3382, 0.5931, 0.6082, 0.6286, and 0.9784 respectively. [17]

4. Conclusion

The market for natural gas in Egypt is influenced by season demand. NGC in Egypt fluctuates from highs in months like Ramadan to lows during the summer vacations.

These season consumptions fluctuate in response to factors of which present Egyptians feasting habits but do not appear to be the only factor.

These consumptions are predicted with greater precision, efficiency gains are realized with the possibility of BI tools utilization for forecasting of natural gas consumption as shown in this paper. Specified consumption is analyzed and appropriate training set, which includes historical consumption data, is defined. Parameters/Coefficients and the goodness of fit of BI tools are obtained using a polynomial fitting, multiple linear regression, and multiple non linear regression models. Analyses of results obtained for training and test sets show that designed/proposed BI models could be useful for NGC forecast problem. The analysis includes comparison between all these methodologies to find out the best model. The challenge was to define the best model, especially in non- linear regression. In this methodology, it is shown that two models with same order may have different goodness of fit. It is well known that, many models can be defined for each order. The challenge was getting the highest goodness of fit with minimum order and the least number of coefficients. In the proposed model more than 97% R2 value is reached.

References

[1] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 2008-12-17.

[2] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.

[3] Spiegel, M. R., 1958 Applied Mathematics for Engineers and Scientists, McGraw Hill, NY

[4] Yalcinoz, T. and Eminoglu, U.: Short term and medium term power distribution load forecasting by neural networks, Energy Conversion and Management, Vol. 46, No. 9-10, pp. 1393-1405, 2005.

[5] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111–117.

[6] Donald Farmer, *Principal Program Manager, US-SQL Analysis Services, Microsoft Corporation,* Smart Business Intelligence Solutions with Microsoft® SQL Server® 2008

[7] M., Wavelet transform and their applications to turbulence, Annu. Rev. Fluid Mech., 1992, 24: 395.

[8] Beccali, M., Cellura, M., Lo Brano, V., Marvuglia, A.: Forecasting daily urban electric load profiles using artificial neural networks, Energy Conversion and Management, Vol. 45, No. 18-19, pp. 2879–2900, 2004.

[9] J. C. Goswami, A. K. Chan, 1999, "Fundamentals of wavelets: theory, algorithms, and applications," John Wiley & Sons, Inc.

[10] I. Daubechies, 1992, "Ten lectures on Wavelets," CBMS-NSF Series in Appl. Math., #61, SIAM, Philadelphia.

[11] C.K. Chui, 1992, “An Introduction to Wavelets,” Academic Press, Boston.

[12] A. Cohen, R. D. Ryan, 1995, “ Wavelets and Multiscale Signal Processing,” Chapman & Hall.

[13] J. J. Benedetto, M.W. Frazier, 1994, “Wavelets-Mathematics and Applications,” CRC Press, Inc.

[14] Chatterjee, S., and A. S. Hadi. "Influential Observations, High Leverage Points, and Outliers in Linear Regression." *Statistical Science*. Vol. 1, 1986, pp. 379–416.

[15] Seber, G. A. F., and C. J. Wild. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience, 2003.

[16] DuMouchel, W. H., and F. L. O'Brien. "Integrating a Robust Option into a Multiple Regression Computing Environment." *Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.

[17] Holland, P. W., and R. E. Welsch. "Robust Regression Using Iteratively Reweighted Least-Squares." *Communications in Statistics: Theory and Methods, A6*, 1977, pp. 813–827.