

## On Forecasting Techniques In Socio-Economic Environment

Mohammed M. N. Quotah

*Associate Professor  
Department of Business Administration,  
Faculty of Economics & Administration,  
King Abdul-Aziz University, Jeddah, Saudi Arabia.*

Mirek Karasek

*Consultant  
Research & Development Department  
Presidency of Civil Aviation, Jeddah, Saudi Arabia.*

( Received 20.1.1418H, and Accepted 13.11.1418H)

ABSTRACT. Several forecasting methods for a target (or response) variable Y is considered. Of these, even simple explanatory-variable-based forecast, let alone multiple explanatory-variable-based causal-chain forecasting models, should provide superior, to most "expert" judgment and various extrapolation-based, time series analyzing methods of Box-Jenkins type and other "black box" related forecasts.

Steps in the development of the relevant theoretical apparatus for judging the quality of data base which, in turn, affects the qualities of individual forecasting methods are described in the text to prove the point.

All the investigated forecasting methodologies have been tested for accuracy and compared with actual data for the forecast period. The data series chosen for the experiment is the annual series "Total International Passengers" at the Jeddah King Abdulaziz International Airport (KAIA) for the years 1975-1987 of which 1975-1981 is considered "historical" data and, based on this interval, forecasts were made for the years 1982 - 1987.

### **The Problem of Forecasting Socio-Economic Phenomena**

A useful term in the understanding of how the process of forecasting relates to what we observe and how it serves, as a means of reporting these observations, is "data credibility" that has to do with our ability to measure and/or sometimes even quantify the socioeconomic data (Shackle, 1968, Morgenstern, 1963, and Alem and Karasek, 1989).

### Credibility of Socioeconomic Data

Credibility of the statistical datum is neither a new term nor a new concept. In fact, there have already been suggestions to start considering a special bi-characteristic datum in economic and other social sciences statistics in the form :

$$(x_i, \phi x_i) \quad (1)$$

This form is favored by Shackle's (1968) work, where  $x_i$  is an  $i$ -th observed and, probably with error, recorded datum. And  $\phi x_i$  is the datum's credibility which can also be defined in terms of the information's inherent error.

We assume that all data are subject to two types of errors:

- a) Random departures  $\epsilon_i$ , from the unknown true value  $x_i$  for which the traditional zero-sum property  $E(\epsilon_i) = 0$  holds; and
- b) Systematic error  $e_i$ , for which the zero-sum property,  $E(e_i) = 0$ , which for reasons explained in literature (Morgenstern, 1963, Streisler, 1970), doesn't enter the realm of realistic expectations in statistics of social sciences.

If we assume  $E(e_i) \neq 0$ , then from literature (Karasek, 1980) it makes sense to express error magnitude in relative term and in its limit values (Mean Absolute Percentage Error MAPE) such that:

$$\phi x_i = [(\sup |e_i|) / (|x_i|)] [100] \quad (2)$$

specially since limit values are generally much easier to estimate or to assess than specific figures; e.g. maximum relative error can be assumed constant for the entire set of data describing a certain socioeconomic phenomenon.

Any quantitative massaging, e.g. Ordinary Least Square (OLS) calculations, of the data where equation (2) prevails, must therefore consider the propagation of an individual data error into the result's total error and we address it this way.

#### Definition 1.

Optimum propagation of an individual data's error (credibility) into the total computational error (credibility) of the model "R" denoted as :

$$\phi x_i \longrightarrow \phi R \quad (3)$$

Is such that it yields minimal  $\phi R$ , where  $\phi R$  is result's alternative of equation (2). We say that the resulting model's error is minimal when all individual data's credibilities in (2) are at their minimum attainable level in the bi-characteristic datum (1), and we employ a data-massaging procedure that yields minimum computational error.

### Graphical Scheme of the Systematic Error in Socioeconomic Data

The inherent error (or credibility)  $\phi x_i$  which according to literature (Dwyer, 1951) is, in socioeconomic data, virtually systematic can be presented graphically in a two dimensional plane seen in Figure (1), where  $\phi x_i$  is the angle between  $x$ -axis and a straight edge  $\phi x_i$  with certain orientation  $\uparrow + \downarrow -$ , and a magnitude.

From literature (Karasek, 1980) and the previous discussion, we postulate important results whose proofs and detailed discussion are found in (Alem and Karasek, 1989, pp.71-74).

**Proposition 1**

If a dependent variable Y is observed with the systematic error portion  $f(\phi Y)$ , then every explanatory variable  $\{X_i\}$  with similarly oriented error coefficient  $f(\phi X_i)$ , where angular orientation (Figure 1) increases the precision of OLS computational results. If the magnitude of the error (or better still 'amplitude') of  $\phi Y = a_1$  and magnitude of any  $\phi x_i = a_i$ , then the regression coefficients' error will be :

$$u = | a_1 - a_i | \tag{4}$$

where the absolute value covers the cases when  $a_i$  is greater than  $a_1$ .

**Proposition 2**

If we change the orientation of the error of a single explanatory variable to contradict the error orientation of the dependent variable, the error-in-result (of the OLS computation), w, increases to :

$$w = a_1 + | a_i | \tag{5}$$

where the absolute value recognizes that the orientation of systematic error is negative.

Both propositions are based on the existence of bi-characteristic data notation as in (1), and suggest their congruency and proximity with a two-dimensional plane. In Figure (1) we see the two important characteristics:

- 1) Amplitude (or magnitude) of systematic error.
- 2) The orientation (or sign) of the systematic error in equation (2) denoted by ↗ +

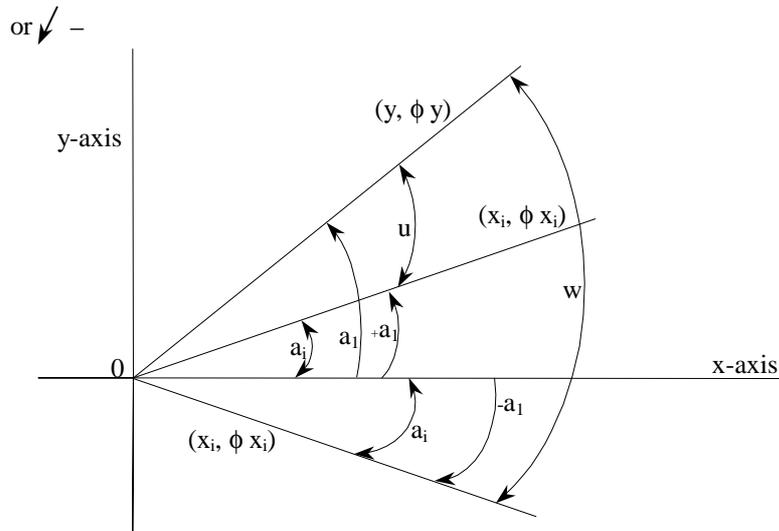


Figure (1)

### Formulating the Problem

Suppose that our goal is to forecast phenomenon  $\dot{Y}$ , henceforth denoted by  $\Psi$ , so that it is distinguished from simple observation  $Y$ ; e.g. as in (1). Suppose also there exists  $m$  different forecasting techniques. Then, using the parallel to equation (2) and to stress the difference between “data observed and data forecasted”, let set  $\Phi Y = d\Psi$ . Our objective is that the best possible forecasting technique will feature :

$$d\Psi_m = \min \quad (6)$$

where  $m$  denotes the  $m$ -th forecasting technique.

### Discussion of Best Forecasting Technique

#### Black Box Systems

Before we start with evaluation and analyses of alternative forecasting methodologies we should start with a definition of the term known in the forecasting community as the “black box systems”.

#### Definition 2:

The so-called “black box” in engineering and systems science is, according to literature (Lilienfeld, 1978) and (Box and Jenkins, 1976) a system that has its well defined ‘inputs’ and ‘outputs’. However, about their interrelation within the system we have no knowledge whatsoever.

#### Causal Chain Systems

As it happens, virtually all to-be-forecast socioeconomic phenomena  $\Psi$  are dependent on, or can be explained by, one or more other variable(s), thus forming the so called “causal chains”, where one or more (usually exogenous) variables  $C_s$  explain either the dependent or target phenomenon (variable)  $\Psi$  directly or explain some middle-layer explanatory variables  $X_i$  first and these, in turn, do the final “explanation”; simplified, so that we look at a system

$$\Psi = f ( X_i ) , \quad (7)$$

$$X_i = F ( X_s ) , \quad (8)$$

while combining equations (7) and (8) we obtain:

$$\Psi = f [ F ( X_s ) ] \quad (9)$$

where  $X_s$  are usually exogenous variables which have an easier-than-the-rest predictability quality or proxies  $X_p$  to be discussed below.

Some of these  $\{ X_i, X_s, X_p \}$  variables can be endogenous ( whose behavior can be affected within the given system that produces target variable  $\Psi$ ). Some, on the other side, can be exogenous ( whose behavior depends on forces and phenomena beyond our control ).

A typical example can be population data series for a country. This target variable  $\Psi$  can be affected both by endogenous cultural, ecological, historically behavioral

aspects, improvements of social and health-care policies, as well as exogenous changes in world economic and political climate and family-size planning devices.

**Forecasting with Causal Chain Models**

We are implicitly assuming here, as it has been clear from the Propositions 1 and 2, that OLS technique is adequate for quantifying the causal chain models of socioeconomic environments, linear regression notwithstanding. Even if relations (7) or (8) are nonlinear we can approximate the nonlinearity by straight edges of linear functions between appropriate inflection points or other important “milestones” (as it is discussed in literature (Karasek, 1990)). The following proposition will be very useful here.

**Proposition 3:**

Assuming that we can find at least one explanatory variable or a proxy (Alem et al 1989, p.39) in system (8) and particularly in system (9) with the same angular orientation AO, (described in proposition 1 and Figure (1) and denoted by superscript ||) as the target variable Y, we say that any forecast  $\Psi_{||}$  is better than  $\Psi$  and, in terms of our objective (6), we write :

$$d\Psi_{||} < d\Psi \tag{10}$$

**Proof :**

Let us take a simple case of relation (8) first, in which both variables have the same angular orientation (AO) ||. Hence,

$$dX_i || < dX_i \quad (\text{whose } X_s \text{ AO are undetermined}) \tag{11}$$

**Then :**

(a) If nothing changed (as far as angular orientation) between  $C_i$  and Y, in system (7) we still obtain resulting target (or dependent) variable, denoted as  $Y_{||}$ , for which the following holds :

$$d\Psi_{||} < d\Psi \tag{12}$$

Simply because from equation (7) and inequality (11) the overall error u in equation (4) as well as w in equation (5) will be smaller.

If in system (7), which is only a simplification of (9), the target variable has the same AO as its explanatory variable(s) or appropriately chosen proxy  $C_p$ , (Alem et al, 1989, pp. 32-36, Maddala, 1992, pp. 464-469), inequality in (10) follows from equation (4).

As to the usefulness of proxies (for dependent variable(s))  $C_p$  in causal chain models forecasting, they should have other features apart from having the same AO as the target variable. They must be, by definition (Barnow, 1976, p.119, Alem et al, 1990, pp.41-44), easier to interpret, have a very strong correlation to the target variable [or at least the same “truth” contents as the target variable or any other variable they stand proxy for] and above all their future trend must be easier to forecast and predict. In fact, many times we wish that the notation of system (9) can be converted to and explained by the well-chosen proxies :

$$\Psi = f [ F( X_s ) ] = f ( X_p ) \tag{13}$$

The most important benefit from causal-chain modeling approach to socioeconomic forecasting is, however, ability to utilize the scenario building (Alem et al, 1990, pp. 89-97) for the forecast trend of the target variable. We select the exogenous and/or proxy variables in equation (13), with the same AO, that have the most significant impact on  $\Psi$  and try several probable trends (scenarios). Then the one closest to the functioning of the realworld environment (from which the exogenous variables are usually drawn) and its assumed trend will also produce the best forecast.

### **Forecasting with “Black Box” Models**

This is a trickier approach because of definition 2. All we can do apart from employing a crystal ball or use some sort of strictly judgmental, for example Delphi-technique, is to analyze more or less exhaustively the historical data of target variable and find a pattern that might serve as the forecast trend.

This set of techniques ranges from simple linear or other naive extrapolation techniques (of moving averages, exponential smoothing, etc.) to Box-Jenkins rather overcomplicated and not very objective (besides only good for short-term forecasting anyway) technique (Armstrong, 1978, pp.158-159). One thing all of these techniques have in common is that they at best propagate but mostly amplify the inherent error into the result.

The most serious shortcomings of these techniques are that by ignoring social and economic environment and its predicted changes, the forecast trends are based solely on the target (dependent) variable’s historical data series. And since we do not know the system that produces the target variable(s), we cannot employ scenarios to align the forecast trend to its natural environment and thus obtain the result closer-to-reality with as minimal as possible systematic error.

### **Example**

Various international aviation sources, such as International Air Traffic Association (IATA), had for years consistently produced medium- and long-term forecasts of the international passenger demand for aviation services in Saudi Arabia that consistently overestimated the actual demand by 20% - 50%. This was achieved by using the perceived future growth trends derived (from the boom years of early and mid 70s) for the international routes in other parts of the world, where the passenger demand has been clearly driven by different factors.

### **Numerical Application Examples**

#### **Data, Variables, and Forecasting Techniques:**

Target variable Y is “Total International Passenger Air Traffic” at the Jeddah’s King Abdulaziz International Airport. The original time series, comes from various issues of PCA (various issues). In the Appendix there are also numerical results of all the investigated forecasting techniques.

The first part of series 1975-1981, is considered to be historical data, as the original model was built in 1981 and published in 1982 (Karasek, 1982). Our aim is to forecast the time interval 1982-1987.





In this model structure, where again the annotated variables are the only relevant ones, the explanatory variable  $X_1$  is missing. However, as in Figure (2), the arrows in the boxes show the direction and degree of relationship between the endogenous, explanatory and exogenous variables. Numbers in the circles of individual arrows represent the elasticity (Barnow, 1976, pp. 416-417).

In so far as variable  $X_{211}$  "Exchange Rate", yields enormous influence on the "Imports" and thus on the fiscal situation of the country, the exchange rate is not tied to the U.S. Dollar (this ratio is pegged) but to "Special Drawing Rights" which represent a basket of major currencies.

**Results**

The following Table (1) summarizes the individual forecasting techniques' MAPE for the forecasting interval 1982 - 1987 based on historical time series 1975-1981.

The individual investigated (or rather tried) forecasting techniques result is listed in succession from the smallest MAPE to the largest one.

Table (1)  
Mean absolute percentage error for different forecasting methods

FORECASTING METHODS	MAPE (1982-1987)
(a) Linear Causal Chain Model.	1.1%
(b) Adaptive Response Rate Exponential Smoothing.	5.7%
(c) Linear Equation ( $Y = a + bX_{11}$ ) "Expatriate".	5.8%
(d) Simple Exponential Smoothing.	8.4%
(e) S-curve Fitting.	10.2%
(f) Linear Exponential Smoothing (best condition=0.2)	13.0%
(g) CONSULTANT 1 (Parsons-Daniel Co.) .	14.9%
(h) CONSULTANT 2 (Hochtief GMBH).	16.9%
(i) Linear Equation: ( $Y = c + dX_1$ ) "GDP".	17.5%
(j) Linear Fitting.	55.9%

Table (2)  
Actual and forecasted data using different forecasting techniques  
(Passenger's data are in \* \* \* \* s)

YEAR	Y	$\Psi_{(a)}$	$\Psi_{(b)}$	$\Psi_{(c)}$	$\Psi_{(d)}$	$\Psi_{(e)}$	$\Psi_{(f)}$	$\Psi_{(g)}$	$\Psi_{(h)}$	$\Psi_{(i)}$
1975	729									
1976	1162									
1977	1796									
1978	2178									
1979	3220									
1980	3499									
1981	3740									
1982	3970	3252	2800	4522	3381	3739	3500	3975	2900	2800
1983	3866	3374	3090	5066	3560	3739	3726	3931	3110	2970
1984	4001	3411	3340	5610	3650	3739	3952	4010	3360	3100
1985	3984	3471	3620	6154	3695	3739	4178	3913	3560	3270
1986	3555	4522	3900	6697	3717	3739	4404	3483	3800	3450
1987	3488	3565	4200	7241	3728	3739	4631	3444	4400	3670

### Discussion of Results

Of the time-series analyzing (i.e. 'black box') forecasting techniques in Table (1), only models (b) and (d) performed well enough to warrant attention.

An interesting and highly discussable result is presented by methodology (i), the linear equation with explanatory variable GDP, which was supposed to yield MAPE superior to any 'black box' result. For that matter, the similar linear equation with explanatory variable "Expatriate Workforce" performed well. The problem with this variable has been discussed in the forecasting techniques.

As for the GDP, being the single most used explanatory variable for (our target variable) "International Passenger Traffic" (Taneja, 1978, pp15-19, 33,43,51), the forecasters (Planning Factors for KAIA, 1983) fell into the trap caused by the specificity of the Kingdom's economic system. In most other countries the GDP characteristics reflect the economic trends reasonably well. In the Kingdom, where the oil (and downstream) revenues are virtually the state revenues, the government has to smooth the sharply (even drastically) changing trend of oil revenues through the monetary (and gold) reserves so that the 'shocks' (on the plus or minus side) would not cause serious economic impact. The final product of these smoothing exercises, based on feedback loops, is GDP. The most important by-product from this notion is that when oil revenues are on the downward trend for long time (as has been the case for the last 8 years or so) the reserves quickly disappear and then the only source of additional financing is bank loans or state bonds (which is the monetary instrument still not much in use in the Kingdom).

To summarize: Gross Domestic Product (GDP), although in most cases quite a useful explanatory variable in its own right, is not any easier to forecast and, particularly in the Kingdom, its systematic error might be very high.

Now, we come to the "corrected" causal chain model, whose evolutionary and already discussed flow-chart is seen in Figure (3).

The final scenario (a) with virtually errorless MAPE at 1.1%, has the entire known world economic trends and long-term projection built-in. It was built on the assumption of constantly subdued oil prices (even when prices go up in current terms the constant unit term does not change upwards) paid in U.S. Dollars (which brought another problem with its drop against other world currencies even at the time after the Gulf War; thereby negated virtually the Kingdom's monopoly in unlimited lifting and bulk selling).

Another assumption of this scenario has acknowledged the overall stagnation of the world economies (of developed countries) which have been preoccupied with post-cold-war economic malaise in Eastern Europe and former USSR.

Another factor in the game has been the slow but identifiable turn to other sources of cheap energy, i.e. building more effective petrol engines, turning Europe from its coal and oil-based heating to natural gas, etc.

We have demonstrated that there is probably no substitution for theoretically well-designed models containing major explanatory (proxy or exogenous) variable(s) whose

future trend(s) can be more easily projected or assumed and those with the same angular orientation (AO) of their systematic errors as that of the dependent variable. At the same time it is necessary to stress that the model complexity is of the least importance. Sometimes, one perfect explanatory (proxy or exogenous) variable, easy to predict and fulfilling contingencies of inequality (10) to (12) and equation (13), works better for given set of data than a more complex model structure.

### References

- Alem, W.** and **M. Karasek**, (1990) Modeling and Forecasting of Air Traffic Demand, *Civil Aviation*, No. 5, PCA-IAP, p.15, Figure (8).
- Alem, W.** and **Karasek, M.**, (1989) *Socioeconomic Modeling and Forecasting in Developing Countries*, The Book Guild Ltd., England.
- Armstrong, J. S.**, (1978) *Long-Range Forecasting*, J. Wiley & Sons, New York, p.158.
- Barnow, B. S.**, (1976) The Use of Proxy Variables when One or Two Independent Variables are measured with Error. *The American Statistician*, Vol. 30, No. 3.
- Box, G.E.P.** and **Jenkins, G. M.**, (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day Inc., San Francisco.
- Dwyer, P.**, (1951) *Linear Computations*, J. Wiley & Sons, p.26.
- Karasek, M.**, (1984) Criterion of Tolerable Limits in Socioeconomic Data Base. *Proc. Amer. Stat. Asso.* (Soc. Stat. Sect.), ASA Press, Washington.
- Karasek, M.**, (1982) Forecasting and Planning the Jeddah Air Traffic with a Mini Model. *J. of Forecasting*, vol. 1, pp. 409-417.
- Karasek, M.**, (1980) Bi-Characteristic Image: A Feasibility Criterion for O L S Treatment of a Cardinal and Ordinal Variables' Mix, *Proc. Amer. Stat. Asso.* (Soc.Stat. Section), ASA Press, Washington.
- Lilienfeld, R.** (1978) *The Rise of Systems Theory*, J. Wiley & Sons, New York.
- Maddala, G. S.** (1992) *Introduction to Econometrics*, Macmillan Pub. Co.
- Morgenstern, O.** (1963) *On the Accuracy of Economic Observations*, Princeton, New York, pp. 4-9.
- PCA Statistical Yearbook**, *PCA-IAP, R & D*, Jeddah, Saudi Arabia.
- Planning Factors for KAIA** (1983): *Air Traffic Forecast 1982-2010*, Parsons & Daniel, June 1983.
- Shackle, G.L.S.** (1968) *Uncertainty in Economics (and Other Reflections)*, Cambridge U.Press, Cambridge.
- Streisler, E. W.** (1970) *Pitfalls in Econometric Forecasting*, The Inst. of Econ. Affairs N-Y, p. 45.
- Taneja, N. K.** (1978) *Airline Traffic Forecasting*, Lexington Books, D.C. Heath & Company, Lexington, Mass.

## أساليب التنبؤ في محيط اقتصادي اجتماعي

محمد محمد نور قوتة

أستاذ مشارك

قسم إدارة الأعمال - كلية الاقتصاد والإدارة

جامعة الملك عبدالعزيز - جدة - المملكة العربية السعودية

ميرك كراسيك

مستشار - إدارة الأبحاث والتطوير

مشاريع المطارات الدولية - الطيران المدني

جدة - المملكة العربية السعودية

**المستخلص:** إن تقدير نموذج التنبؤ لأي بيانات مهمة صعبة ومعقدة نظرا للحاجة إلى الخبرة والقدرة علي معرفة النموذج المناسب للحالة الراهنة. ولقد كان الهدف من هذا البحث هو تطبيق عدة أساليب من نماذج التنبؤ للمتغير ص والذي يعتبر بمثابة الاستجابة response أو الهدف target وكان من ضمن هذه النماذج النموذج البسيط والذي يعتمد علي متغير واحد، بالإضافة إلى النماذج التي تعتمد علي متغيرات اقتصادية عديدة. ولا شك أن استخدام أي من هذه النماذج في عملية التنبؤ يعتبر أفضل بكثير من استخدام الحكم الشخصي والخبرة وحدها. ولقد بدأ البحث في تطوير أساليب ونظريات للحكم على جودة قاعدة البيانات المستخدمة في عملية التنبؤ، والتي بدون أدني شك لها التأثير الكبير علي دقة نتائج النموذج المستخدم لعملية التنبؤ، ولرؤية هذه الحقيقة فقد تم مناقشة ذلك في البحث لإثبات هذه الفكرة.

ولقد تم فحص النماذج (investigated) المستخدمة من حيث الدقة في التنبؤ وذلك بمقارنة نتائج النماذج بالبيانات الحقيقية (الفعلية) لفترة التنبؤ التي تم اعتبارها. على أن مجموعة البيانات التي تم اختيارها لهذه التجربة هي بيانات سنوية مستخرجة "بمجموع المسافرين دوليا" عن طريق مطار الملك عبد العزيز بجدة للفترة من سنة ١٩٧٥ م إلى سنة ١٩٨٧ م والتي تعتبر بيانات تاريخية "واقعية" وبناء على هذه الفترة المستخدمة تم القيام بعملية التنبؤ للفترة عن ١٩٨٢ م إلى سنة ١٩٨٧ م.