FRACTIONALLY-INTEGRATED AND MARKOV-SWITCHING MODELS TO FORECAST THE DEMAND OF A FAST-GROWING EMERGING AIR TRANSPORTATION MARKET

Diego Javier Gonzales Vega
Alessandro Vinícius Marques de Oliveira

Instituto Tecnológico de Aeronáutica - ITA

E-mails: nectar@ita.br, avmoliveira@gmail.com

ABSTRACT

This paper presents a methodology using a portfolio of time-series models, from conventional ARMA to a regime-changing framework. The objective is to develop an air transportation demand modeling to inspect potential structural breaks in the Brazilian market, due to solve underlying issues of new demand creation. Out-of-sample forecasting is used to generate comparative metrics aiming at both selection and validation of what we call a "champion model". Results indicate better performance of more complex models such as the fractionally integrated and the Markov-switching models. Ex ante knowledge of the interaction of structural breaks and unit root can prove useful in the modeling analysis. The demand forecast of the champion model is in line with the recent accelerated growth of the Brazilian air transportation market, roughly 7% per year.

Keywords: Air transportation demand, time-series modeling, regime-change analysis, structural break vs. unit root, forecast.

1. INTRODUCTION

Due to the development of new consumers and the increase in frequency of travel from existing consumers, air transportation demand in the world have been growing at an accelerated rate. In Brazil, this phenomena has modified the behavior of the demand, at rates never seen before, creating a natural experiment that allow us to model the demand in new ways. The main goal of this paper is to present a methodology for modeling and choosing a forecast model of air transport demand, using a wide portfolio of econometric time-series models, –in contrast with the several works that show a previously chosen model– using out-of-sample performances metrics of long term forecast. As a secondary objective, testing structural break of consumer behavior using aggregated data will accomplish a separation between modeling characteristics, being those, models of one and multiple regimes. Also to compare the results to other works in the area and present a forecast of air transportation demand up to 2014.

The main motivation of the research is to present this sequence of time-series models of air transportation demand to achieve robust models for the sample, this models lead us to compare the out-of-sample forecast fitness metrics to aid us in the selection of better explanatory model of the sample, which is also validated by another sample within the available data; afterward this model receive the name of "champion model", that would accomplish accurate forecasts for airport planning and investments in the Brazilian market, given the proximity of important events such as the World Cup, the Olympic Games, and further planning (privatizations, expansions, etc.). To achieve this, alternative models to estimate air transportation demand were studied to take advantage of its unique features, and the treatment they give to data.

Often within any given data a sample is chosen and the modeling process begin but when in the presence of structural breaks, there is an apparent change in the parameters of the model which leads us to explain the behavior of the model as a sum of many sub-models, one for each segment before and after those structural breaks, these will be called “regimes”. Hence the regime analyses of some types of time-series models should allow us to prevent these issues. To achieve an accurate forecast it’s necessary to properly model the demand. This study will demonstrate the steps for an efficient choice of model, beginning for the necessary conditions of consistency and the sufficient conditions of stability in which structural breaks analysis is very important in order to generate forecasts.

Very often the literature about air transportation demand begin their analysis from a previously chosen model, a model that embodies the preferred characteristics to behave in a given scenario, like Karlafitis (2008) or Profillidis (2000). As this paper have a more academic approach, the intent is to show that process as a methodology, for any author to be able to dive into that selective process; even if some models outperform others, the choice depend not only in the results of the forecast, but in the robustness of its assumptions.

Multiple econometric software like EViews®, JMulti® and OXMetrics® were used for modeling and forecast the demand; the database was obtained through the Brazilian Central Bank (BCB), the Brazilian Institute of Geopolitics and Statistics (IBGE), and the National Agency of Civil Aviation (ANAC) website. The variables are the domestic revenue passenger kilometer (RPK), the gross domestic production of Brazil (GDPBR) and the Yield (revenue received for carrying a passenger); Additional secondary variables were used to aid in the forecast of exogenous variables like: gross domestic production of the USA (GDPUS), the cost of aviation fuel kerosene (QAV), and exchange rates R/$ (USD). The data range is from January of 2002 to December of 2010.

In this paper we employ the following steps in order to reach a “champion model” for forecasting air transportation demand:

1st) Run all models for the sample 2002 to 2008, ensuring that the estimated parameters pass the requisites of stability and significance (individual and global). We use one regime models - causal models, ARMA (Autoregressive moving average models) and simultaneous equation models as the vector autoregressive (VAR) and a vector of error correction (VEC).

2nd) Test presence of structural break and/or unit root in the principal variables, with the intention to indicate the utility of such analysis and to create a bridge between regime related groups, then test the multiple regime models:

- Smooth transition models.- This model use a transition function (which can be logistic or exponential) to show the changes between regimes, this function use a known transition variable. The computational process is limited to analyze just two regimes. The literature is comprised in the work of Teräsvirta T. (1994) and van Dijk, D. and P.H. Franses (1999).
- Markov-switching models.- The main difference of this models is that the transition variable is hidden in the form of a probabilistic function known as “Markov-chain”, also it accepts switching intercept and parameters between multiple regimes. Economic application of the literature of this model begins with Hamilton J. (1989) and was further worked by Kim J. (1994).

3rd) Then calculate the MAPE (Mean absolute percentage error), and RMSE (Root mean square error) of the Out-of-sample forecast.

4th) Afterwards, we validate the selection reworking step c with other samples, thus selecting a champion model and present the forecasts.

After a brief introduction in the chapter one, the chapter two will present the literature review concerning time-series methods of model air transportation demand. The review includes a resume of the econometric and statistic theory, a brief look at the development of such methods by several authors and how that specific theory was used to model air transportation demand. It was also presented the actual work in the area of forecast of air transportation demand in Brazil and the world, finalizing with a review of issues regarding how to deal with structural break in general and in the modeling. The chapter three has a presentation of the steps used to model the demand, a description of the variables, the parameters studied, the metrics taken in consideration to discriminate the robustness of each model and an explanation of the assumptions for the forecast of the exogenous data. Chapter four will show the modeling, the estimation and the out–of–sample forecast of all the models concluding with a resume of the consistency, significance and stability of the parameters. A resume of the metrics used to select and validate the champion model and a forecast to 2014. In chapter five we conclude the work showing how the fractioned integrated ARMA out-performs simultaneous equation model and how the multiple-regime models, which have close accuracy and offers a better explanation to structural breaks. Also presenting recommendations for future work, i.e. take into account the break analysis over unit root, and to study more the scenarios over static comparative.

2. TIME SERIES MODELING METHODOLOGIES AND THE EMPIRICAL APPROACH

This section will present a brief resume of the extensive work of the time-series models applied in the modeling, highlighting the origins of the methodology.

2.1.Time-Series Models

2.1.1. Linear Models

In the field of economics, numerous hypotheses and theories have been proposed in order to describe the behavior of economic agents and the relationships between economic variables. Although these propositions may be theoretically appealing and logically correct, they may not be practically relevant unless they are supported by real world data.

A theory with supporting empirical evidence is of course more convincing. Therefore, empirical analysis has become an indispensable ingredient of contemporary economic research. By econometrics we mean the collection of statistical and mathematical methods that utilize data to analyze the relationships between economic variables, Davidson and MacKinnon (1993).

A leading approach in econometrics is the regression analysis. For this analysis one must first specify a regression model that characterizes the relationship of economic variables, Greene (2000). The simplest and most commonly used specification is the linear model, Goldberger (1991). The linear regression analysis then involves estimating unknown parameters of this specification, testing various economic and econometric hypotheses, and drawing inferences from the testing results. One of the most important estimation methods in linear regression is the ordinary least squares (OLS).

Suppose that there is a variable "*y*", whose behavior over time (or across individual units) is of interest to us. A theory may suggest that the behavior of *y* can be well characterized by some function "*f*" of the variables $x_1, \ldots, x_k$. Then, $f(x_1, \ldots, x_k)$ may be viewed as a "systematic" component of y provided that no other variables can further account for the behavior of the residual $y - f(x_1, \ldots, x_k)$. In the context of linear regression, the function f is specified as a linear function. The unknown linear weights (parameters) of the linear specification can then be determined using the OLS method, Harvey (1990), Intriligator et al. (1996), Johnston (1984), Judge et al. (1988), Maddala (1992), Ruud (2000), and Theil (1971), among many others..

There is a special case of linear modeling that is frequently used in economics, the log-linear transformation of the Cobb-Douglas function, journal American Economic Review (1928):

$$Y_T = aX_{1t}^{\beta 1}X_{2t}^{\beta 2} \rightarrow logY_T = \alpha_0 + \beta_1 logX_{1t} + \beta_2 logX_{2t} + \varepsilon_t \qquad (1)$$

Mostly used to study production, can also be used at demand modeling, the model (1) can represent the scale of the market and one of its perks allow us to study the elasticities of the variables as the estimated parameters, we use this simplicity as our initial model.

2.1.2. ARMA Modeling

The ARMA-based are sophisticated univariate time-series forecasting methods, which are sometimes referred to as time-series analysis in the time domain complementary to methods such as ''spectral analysis,'' which decompose the time series into component cycles of varying frequency, which are referred to as time-series analyses in the frequency domain (Gottman, 1981).

Analyses in these respective domains have complementary strengths and weaknesses (although the two approaches are in one sense the same because parameters derived from analyses in one domain can be mathematically related to those from the other). Methods in the frequency domain are better suited to searching for physically or biologically based periodicities in time series, although ARMA also has a limited capability to detect periodicities.

ARMA-based methods are better suited to assessing effects associated with non-cyclical experimental paradigms while removing autocorrelation and confounding temporal trends unrelated to experimental conditions, Chu (2009).

A time series Y_t performs as an ARMA(p,q) process if its stationary and if for every t:

$$Y_t - \varphi_1 Y_{t-1} - \cdots - \varphi_p Y_{t-p} = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}; \qquad (2)$$

Where: $\{e_t\} \sim N(0, \sigma^2)$

Using the Backshift operator we can write the equation (2) concisely as:

$$\varphi(B)Y_t = \theta(B)e_t$$

Where φ(B) and θ(B) are respectively the regressive operator and the moving average operator, both polynomial in B.

Note that when φ(B) = 1 then ARMA(p,q) is equivalent to MA(q) and when θ(B) = 1 then ARMA(p,q) is equivalent to AR(p). Such processes are often denoted as ARMA(0,q) and ARMA(p,0) to stress the fact that the moving average model and the autoregressive model are members of the ARMA models family.

A generalization of ARMA models which incorporates a wide class of nonstationary time series is obtained by introducing the differencing into the model. The simplest example of a nonstationary process which reduces to a stationary one after differencing is Random Walk. A

Random Walk is a nonstationary AR(1) process with the value of the parameter φ equal to 1, that is the model given by:

$$Y_t = \varphi Y_{t-1} + e_t; \quad \text{Where: } \{e_t\} \sim N(0, \sigma^2)$$

Its autocovariances depend on time as well as on lag. However, the first difference

$$\nabla Y_t = Y_t - Y_{t-1} = e_t;$$

Is a stationary process, as it is just the white noise e_t. So, with Normal distribution in the ARMA class then ∇Y_t is an ARMA(0,0) process, or in ARIMA notation it is ARIMA(0,1,0) process as it is obtained after first order differencing of Y_t. A process Y_t is said to follow an Integrated ARMA model, denoted by ARIMA(p,d,q), if:

$$\nabla_d Y_t = (1 - B)^d Y_t;$$

Is ARMA(p,q). We write the model as

$$\varphi(B)(1 - B)^d Y_t = \theta(B) e_t; \quad \{e_t\} \sim N(0, \sigma^2)$$

Where d is a positive integer that controls the level of differencing (or, if d = 0, this model is equivalent to an ARMA model). Conversely, applying term-by-term differencing d times to an ARIMA (p, d, q) process gives an ARMA (p, q) process.

Sometimes a seasonal effect is suspected in the model. In this case it is often considered better to use a seasonal ARIMA (SARIMA) model than to increase the order of the AR or MA parts of the model. Let us assume that there is seasonality in the data, but no trend. Then we could model the data as:

$$V_t = s_t + Y_t$$

Where Y_t is a stationary process. The seasonality component is such that:

$$s_t = s_{t-h}$$

Where h denotes the length of the period, using the approach of Chatfield (2004) it is possible to remove the seasonal effect from the data by differencing by lag h. Thus introducing the lag-h operator:

$$\nabla_h V_t = V_t - V_{t-h} = V_t - B^h V_t = (1 - B^h) V_t$$

Which gives:

$$\nabla_h V_t = s_t + Y_t - s_{t-h} - Y_{t-h} = \nabla_h Y_t;$$

Hence, this operation removes the seasonality effect. This fact leads to introducing the seasonal ARMA model, denoted by $ARMA(P,Q)_h$, which is of the form

$$\varphi(B^h) X_t = \theta(B^h) e_t$$

Where:

$$\varphi(B^h) V_t = (1 - \varphi_1 B^h - \varphi_2 B^{2h} - \cdots - \varphi_P B^{Ph}) V_t$$

And:

$$\theta(B^h) e_t = (1 - \theta_1 B^h - \theta_2 B^{2h} - \cdots - \theta_Q B^{Qh}) e_t$$

Are, respectively, the seasonal AR operator and the seasonal MA operator, with seasonal period of length h.

When we combine seasonal and non-seasonal operators we obtain a model:

$$\Phi(B^h)\varphi(B)Y_t = \Theta(B^h)\theta(B)e_t$$

Mixed seasonal ARMA is a stationary process. In practice however we often have nonstationary processes. Seasonal nonstationarity can occur when the process is nearly periodic in the season and the seasonal component varies slowly from period to period (say from year to year) according to a random walk.

This leads to a very general seasonal autoregressive integrated moving average (SARIMA) model written as follows:

$$\Phi(B^h)\varphi(B)\nabla_H^D\nabla^d Y_t = \alpha + \Theta\left(B^h\right)\theta(B)e_t$$

And denoted by SARIMA(p,d,q) x $(P,D,Q)_h$

Since the work of Box and Jenkins (1970), the short-range dependent ARIMA process has become popular in empirical studies on time series. As a generalization of this type of model that incorporates long-range dependence, Granger and Joyeux (1980) and Hosking (1981) independently discuss fractionally integrated processes in which the difference parameter "d" is allowed to be a non-integer. Their proposal of fractional integration at that time has no physical but only a mathematical sense. However, such models have been analyzed by statisticians in the physical sciences since at least 1950.

According to McLeod and Hipel (1978), a second way of characterizing long memory is in the frequency domain: a process is said to have long memory if the spectral density function has a pole or singularity at zero frequency. Then, a popular model satisfying these two properties is the fractionally integrated one. So, in some cases, the time series is neither consistent with an I(1) process nor an I(0) process. This means that the autocorrelations suggest that the original series in general appears not to be stationary, and the first difference is over-differenced. If a series reveals long memory, there is a persistent temporal dependence even between distant observations.

The ARFIMA (p, d, q) model is represented by:

$$\varphi(B)(1-B)^d Y_t = \varphi(B)e_t$$

But the difference operator $(1 - B)$ is raised to a fractional power d, denoting the fractional order of integration. By examining the Wold decomposition and the autocorrelation coefficients, it is possible to show that they have a very slow rate of hyperbolic decay. This hyperbolic decay at high lags distinguishes the series with long memory from the series with short memory, and is the main characteristic of the empirical identification. For [$-0.5 < d < 0.5$], the process is covariance stationary, while [$d < 1$] implies mean reversion (after a shock, the series tends to revert to its mean level). When $d = 0$, the model reduces to the classical ARIMA model (Box and Jenkins, 1970). For [$0.5 < d < 0$], the ARFIMA process is said to be 'anti-persistent' (Mandelbrot, 1977) or to have 'intermediate memory', and all of its autocorrelations (excluding lag zero) are negative and decay hyperbolically to zero.

The ARFIMA(p, d, q) separately and flexibly describes both long- and short-term behavior. The parameter d determines the long-memory aspects of the process, i.e., the long-term correlations and the behavior of the spectral density near zero frequency. On the other hand, the AR and MA parameters provide away to model the short-term correlations and the spectral density for frequencies not near zero, independently of the long-term description provided by d.

In our study we are including exogenous variables to the modeling, hence the ARMA models become ARMAX, the convenience of this comes from the fact that we want to study the relevance of the GDP and YIELD on the RPK aside from the relevance of the historical behavior of the variable. Given the special complexity of the ARFIMAX models we will use a different approach to estimate the parameters for that we use the OxMetrics® non linear OLS; the other ARMAX models will be estimated with EViews® linear maximum likelihood.

2.1.3. Simultaneous Equation Models

This models can and will sustain exogenous variables, in this case the GDPBR will serve as one, while the YIELD serve us as the other endogenous variable, the selected variables must have economic influences on each other. In other terms, there must be causality between them. The overparameterization and loss of degrees of freedom problems must be avoided to capture the important information in the system. An additional case when three variables assume endogenous behavior is studied with the intention to show that there shouldn't be a gain/loss of estimating power by adding equations to the model.

The vector autoregression (VAR) model is used for analyzing the interrelation of time series and the dynamic impacts of random disturbances (or innovations) on the system of variables. Following Enders (1995), consider a simple bivariate first order VAR, i.e. VAR(1) model:

$$y_t = \beta_{10} - \beta_{12}x_t + \alpha_{11}y_{t-1} + \alpha_{12}x_{t-1} + \mu_{yt} \; ; \quad (3)$$

$$x_t = \beta_{20} - \beta_{21}y_t + \alpha_{21}y_{t-1} + \alpha_{22}x_{t-1} + \mu_{xt} \; ; \quad (4)$$

Where it is assumed that both y_t and x_t are stationary; μ_{yt} and μ_{xt} are white noise with standard deviations of σ_y and σ_x, respectively; μ_{yt} and μ_{xt} are uncorrelated.

Equations (3) and (4) constitute a two variable first order VAR model. In this system y_t is influenced by current and past values of x_t, and x_t is influenced by current and past values of y_t.

Thus the VAR(1) model captures the feedback effects allowing current and past values of the variables in the system. The coefficients β_{12} and β_{21} represent the contemporaneous effects of a unit change of x_t on y_t and of y_t on x_t, respectively;

α_{12} is the effect of a unit change of x_{t-1} on y_t,

α_{21} is the effect of a unit change of y_{t-1} on x_t.

Hence y_t and x_t have mutually contemporaneous effects on each other in the system. The disturbance terms μ_{yt} and μ_{xt} are shocks or innovations in y_t and x_t. The term μ_{yt} has an indirect contemporaneous influence on x_t if $\beta_{21} \neq 0$, and μ_{xt} has an indirect contemporaneous effect on y_t if $\beta_{12} \neq 0$.

Equations (3) and (4) represent the structural VAR model. This model uses economic theory to describe the dynamic relationship between variables. However, appearance of the endogenous variables on both sides of the equations complicates the estimation and inference processes. A standard VAR model can be applied to overcome the difficulties of the structural VAR model - Sims (1980). The standard form of VAR model for the two variable cases can be written as:

$$y_t = \gamma_{10} + \gamma_{11}y_{t-1} + \gamma_{12}x_{t-1} + \varepsilon_{1t} \; ; \quad (5)$$

$$x_t = \gamma_{20} + \gamma_{21}y_{t-1} + \gamma_{22}x_{t-1} + \varepsilon_{2t} \; ; \quad (6)$$

In (5) and (6), the terms ε_{1t} and ε_{2t} are random innovations or shocks, and they are correlated if there are contemporaneous effects of y_t on x_t and of x_t on y_t, but the terms ε_{1t} and ε_{2t} are uncorrelated if there are not contemporaneous effects on each other. In the system each endogenous variable is determined by a function of the lagged values of the two endogenous variables. The OLS is the appropriate method since only lagged variables are included on the right hand side of the each equation, and also disturbances are assumed to be serially uncorrelated with constant variance.

The included variables in a VAR model are selected according to the relevant economic theory. The appropriate lag length must be determined by allowing a different lag length for each equation at each time and choosing the model with the lowest AIC and SBC values. The same sample period must be considered for different lag lengths. If the lag length is too small, the model will be misspecified; if it is too large, the degrees of freedom will be lost.

The VAR analysis determines the interrelationship among the economic time series rather than the parameter estimates. The residual correlation in the VAR model reveals the interaction of the variables in the previous periods.

The main uses of the VAR model are the impulse response analysis, variance decomposition, and Granger causality tests. An impulse response function traces the response of the endogenous variables to one standard deviation shock or change to one of the disturbance terms in the system. A shock to a variable is transmitted to all of the endogenous variables through the dynamic structure of the VAR. Therefore, an impulse response function shows the interaction between/among the endogenous variables sequence.

Variance decomposition analysis provides information about the dynamic behavior of the model and the relative importance of each random disturbances or innovation in the VAR. Variance decomposition shows the proportion of the movements in the endogenous variable sequence as a result of its own shocks against Shocks to other variables.

VAR models are used to test the causality relationship between the variables in the system. Granger causality provides important information about the exogeneity, in other words x_t is defined as an exogenous variable if the current and past values of y_t do not affect x_t. In that case, all the coefficients on current and past y_t are zero. Granger noncausality shows that x_t sequence is independent of both the u_{yt} shocks and yt sequence.

The Vector Error Correction (VEC) Model

Engle and Granger (1987) point out that a linear combination of two or more nonstationary series may be stationary. The stationary combination may be interpreted as the cointegration, or equilibrium relationship between the variables.

For example, in our case the demand model, if the RPK and Yield are cointegrated, then there exists a long run relationship between them.

However, if they are not cointegrated, then consumption might drift above or below yield in the long run, implying that consumers either spend too much or increase savings. A VEC model is a restricted VAR model. The VEC specification restricts the long run behaviour of the endogenous variables to converge to their long run equilibrium relationships and allow the short run dynamics.

Consider the relationship between RPK and YIELD in a simple EC model:

$$\Delta RPK_t = \pi_1(RPK_{t-1} - \delta YIELD_{t-1}) + \mu_{1t}, \pi_1 > 0 \; ; \qquad (7)$$

$$\Delta YIELD_t = -\pi_2(RPK_{t-1} - \delta YIELD_{t-1}) + \mu_{2t}, \pi_2 > 0 \; ; \qquad (8)$$

Where μ_{1t} and μ_{2t} are white noise disturbances, π_1 and π_2 represent the speed of adjustment parameters. π_1, π_2 and δ are the positive parameters.

The cointegrating term (RPK_{t-1} - $\delta YIELD_{t-1}$) is the error correction term since the deviation from long run equilibrium is corrected gradually through short run adjustments. RPKt and $YIELD_t$ are the two endogenous variables.

In an EC model, the short run dynamics of the variables in a system are influenced by the deviations from the long run equilibrium. For example, RPK_t and $YIELD_t$ change in response to the previous period's deviation from long run equilibrium.

In the VEC model if:

- The deviations are positive, i.e.(RPK_{t-1} - $\delta YIELD_{t-1}$) >0, then the yield would rise and the rpk should fall, as the other things are constant. Long run equilibrium is achieved as $RPK_{t-1} = \delta YIELD_{t-1}$.
- $RPK_{t-1} = \delta YIELD_{t-1}$, then RPK_t and $YIELD_t$ change only in response to u_{1t} and u_{2t} shocks.
- π_1 is large, then RPK_t shows greater response to the previous period's deviation from long run equilibrium.
- π_1 is small, then RPK_t is unresponsive to the previous period's deviations from equilibrium.
- π_2 =0, then $YIELD_t$ changes only in response to μ_{2t}, since $\Delta YIELD_t = \mu_{2t}$. Hence RPK_t changes to eliminate any deviations from long run equilibrium.
- π_1 =0 or π_2 =0, there would not be a causality relationship between cointegrating variables.
- π_1 =0 and π_2 =0, there would not be a long run equilibrium relationship between the two variables. The VEC or cointegration models cannot be used for these variables.

The crucial point of using VEC models is the requirement of cointegration between the two variables with the cointegrating vector (1 - δ). In other words, (RPK_{t-1} - $\delta YIELD_{t-1}$) must be stationary.

2.1.4. Smooth Transition Models

Regime-switching models are time-series models in which parameters are allowed to take on different values in each of some fixed number of "regimes." A stochastic process assumed to have generated the regime shifts is included as part of the model, which allows for model-based forecasts that incorporate the possibility of future regime shifts. In certain special situations the regime in operation at any point in time is directly observable. More generally the regime is unobserved, and the researcher must conduct inference about which regime the process was in at past points in time. The primary use of these models in the applied econometrics literature has been to describe changes in the dynamic behavior of macroeconomic and financial time series.

Regime-switching models can be usefully divided into two categories, "threshold" models and "Markov-switching" models. The primary difference between these approaches is in how the evolution of the state process is modeled. Threshold models, introduced by Tong (1983), assume that regime shifts are triggered by the level of observed variables in relation to an unobserved threshold.

In the work of Teräsvirta (2000) we observe the smooth transition autoregressive STAR model for a univariate time series y_t is given by:

$$y_t = \varpi' z_t\big(1 - G(s_t, \tau, c)\big) + \omega' z_t G(s_t, \tau, c) + u_t\ ; \qquad (9)$$

Where $z_t = (y_t', x_t')'$ is an $((m + 1) \times 1)$ vector of explanatory variables with $yt' = (1, y_{t-1}, \ldots, y_{t-p})'$ and $x_t\ 0 = (x_{1t}, \ldots, x_{kt})'$. ϖ and ω are the parameter vectors of the linear and the nonlinear part respectively. The transition function $G(s_t,\tau,c)$ depends on the transition variable s_t, the slope parameter τ and the vector of location parameters c.

Two interpretations of the STAR model are possible. On the one hand, the STAR model be thought of as a regime-switching model that allows for two regimes, associated with the extreme values of the transition function, $G(s_t,\tau,c) = 0$ and $G(s_t,\tau,c) = 1$, where the transition from one regime to the other is smooth. On the other hand, the STAR model can be said to allow for a 'continuum' of regimes, each associated with a different value of $G(s_t,\tau,c)$ between 0 and 1. In this paper we will use the 'two-regime' interpretation.

The regime that occurs at time t can be determined by the observable variable s_t and the associated value of $G(s_t,\tau,c)$. Different choices for the transition function $G(s_t,\tau,c)$ give rise to different types of regime-switching behavior. A popular choice $G(s_t,\tau,c)$ is the first-order logistic function:

$$G(s_t, \gamma, c) = \frac{1}{\left(1 + e^{-\tau(s_t - c)}\right)} \quad ; \tau > 0 \quad (10)$$

And the resultant model is called the logistic STAR [LSTAR] model. The parameter c can be interpreted as the threshold between the two regimes, in the sense that the logistic function changes monotonically from 0 to 1 as s_t increases.

The parameter determines the smoothness of the change in the value of the logistic function and, thus, the smoothness of the transition from one regime to the other. As becomes very large, the change of $G(s_t,\tau,c)$ from 0 to 1 becomes almost instantaneous at $s_t = c$ hence, the LSTAR model with (11) nests a two-regime threshold autoregressive [TAR] model as a special case. In case $s_t = y_{t-d}$, this model is called a self-exciting TAR [SETAR] model.

The other possible choice comes from the exponential function:

$$G(st, \gamma, c) = 1 - e^{-\tau(s_t - c)^2} \qquad ; \tau > 0 \qquad (11)$$

The exponential function has the property that $G(s_t,\tau,c)$ tends to 1 both as st tends to infinite positive and negative whereas $G(s_t,\tau,c) = 0$ for $s_t = c$. The resultant exponential STAR [ESTAR] model has been applied to real exchange rates by Michael, Nobay and Peel (1997) and Taylor, Peel and Sarno (2000) and to real effective exchange rates by Sarantis (1999).

A drawback of the exponential function is that for either $\tau \to 0$ or $\tau \to 1$, the function collapses to a constant (equal to 0 and 1, respectively). Hence, the model becomes linear in both cases and the ESTAR model does not nest a SETAR model as a special case. If this is thought to be desirable, one can instead use the second-order logistic function:

$$G(s_t, \gamma, c) = \frac{1}{\left(1 + e^{-\tau(s_t - c_1)(s_t - c_2)}\right)} \quad ; c_1 \leq c_2,\ \tau > 0 \qquad (12)$$

Where now $c = (c_1; c_2)'$, as proposed by Jansen and Teräsvirta (1996). In this case, if $\tau \to 0$, the model becomes linear, whereas if $\tau \to 1$and $c_1 \neq c_2$, the function $G(s_t,\tau,c)$ is equal to 1 for $s_t < c_1$ and $s_t > c_2$ and equal to 0 in between. Hence, the STAR model with (12) nests a restricted three-regime (SE)TAR model, where the restriction is that the outer regimes are identical.

To estimate this model we use the econometric software JMulti®, it can identify the type of transition function (10) or (12) by testing presence of non-linearity, in a different step estimate the smooth and slope parameters (γ and c), and the transition parameters of the linear and non linear part. While studies show that is possible to represent a multiple-regime STAR, the most literature is center on the two-regime approach, for the multiple regime we use the Markov-switching models.

2.1.5. Markov-Switching Models

Markov-switching models, introduced to econometrics by Goldfeld and Quandt (1973), Cosslett and Lee (1985), and Hamilton (1989), assume that the regime shifts evolve according to a Markov chain.

Given the easiness of how this methodology interacts with macroeconomic variables and its fluctuation overtime by international or political shocks they became really popular as a modeling tool for regime-switching models of measures of economic output, such as real Gross Domestic Product (GDP), used to model and identify the phases of the business cycle.

Time series in the economic field often exhibit breaks in their behavior, associated with events such as changes in government policy or international law. Of particular interest to economists is the apparent tendency of many economic variables to behave distinctly during economic downturns, when external shocks rather than their long-run tendency to grow govern its economic dynamics. Sudden changes are also a common feature of aggregated transportation data, so we use an approach described often to study financial theoretical calculations for how abrupt changes in fundamentals should show up in asset prices (Ang and Bekaert, 2003; Garcia, Luger, and Renault, 2003; Dai, Singleton, and Wei, 2003).

Suppose that the typical historical behavior could be described with a first-order autoregression:

$$y_t = \alpha_1 + \varphi_1 y_{t-1} + \varepsilon_t \; ; \qquad (13)$$

With $\varepsilon_t \sim N(0, \sigma^2)$, which seemed to adequately describe the observed data for t = 1, 2, ..., t_0.

Suppose that at date t_0 there was a significant change in the average level of the series, so that we would instead wish to describe the data according to:

$$y_t = \alpha_2 + \varphi_1 y_{t-1} + \varepsilon_t \; ; \qquad (14)$$

For $t = t_0 + 1, t_0 + 2, ...$ This fix of changing the value of the intercept from α_1 to α_2 might help the model to get back on track with better forecasts, also correcting structural break, but it is rather insufficient an interpretation that such occurrence could have generated the data. We surely would not want to affirm a deterministic change from α_1 to α_2 at date t_0 as an event that anyone would have been able to predict with certainty looking ahead from date t = 1. Instead there must have been some imperfectly predictable forces that produced the change. Hence, rather than claim that expression (13) governed the data up to date t_0 and (14) after that date, what we must have in mind is that there is some larger model encompassing them both:

$$y_t = \alpha_{st} + \varphi_1 y_{t-1} + \varepsilon_t \; ; \qquad (15)$$

Where s_t is a random variable that, as a result of institutional changes, happened in our sample to assume the value $s_t = 1$ for t = 1, 2, ..., t_0 and $s_t = 2$ for $t = t_0 + 1, t_0 + 2, \ldots, T$. A complete description of the probability law governing the observed data would then require a probabilistic model of what caused the change from $s_t = 1$ to $s_t = 2$. The simplest such specification is that s_t is the realization of a two-state Markov chain with

$$\Pr(s_t = j | s_{t-1} = i, s_{t-2} = k, ..., y_{t-1}, y_{t-2}, ...) = \Pr(s_t = j | s_{t-1} = i) = p_{ij} \quad (16)$$

Assuming that we do not observe s_t directly, but only infer its operation through the observed behavior of y_t, the parameters necessary to fully describe the probability law governing y_t are then the variance of the Gaussian innovation σ^2, the autoregressive coefficient φ_1, the two intercepts α_1 and α_2, and the two state transition probabilities, p_{11} and p_{22}.

The specification in (16) assumes that the probability of a change in regime depends on the past only through the value of the most recent regime, though, nothing in the approach described below precludes looking at more general probabilistic specifications. But the simple time-invariant Markov chain (16) seems the natural starting point and is clearly preferable to acting as if the shift from α_1 to α_2 was a deterministic event. Permanence of the shift would be represented by $p_{22} = 1$, though the Markov formulation invites the more general possibility that $p_{22} < 1$. Certainly in the case of business cycles or financial crises, we know that the situation, though dramatic, is not permanent. Furthermore, if the regime change reflects a fundamental change in demand behavior, the prudent assumption would seem to be to allow the possibility for it to change back again, suggesting that $p_{22} < 1$ is often a more natural formulation for thinking about changes in regime than $p_{22} = 1$.

This type of modeling allow us to switch not only the intercept but also the ARMA components and the variance, the software used for this modeling is the OxMetrics®, it can, by the same process that detect the regime probability, give forecast for each regime and establish the most probable regime forecast.

2.2. Air Transportation Demand Forecasting: Brazil and The World

Carson (2010) studies forecast of air transportation demand analyzing the convenience of aggregated forecast versus the sum of individual airports forecast under the hypothesis that the heterogeneity of behavior in each region would add accuracy to the aggregated forecast, his results show exactly that, but also that the necessity of so many data made the work extenuate, in the Brazilian case there are few data of regional airports which difficult that heterogeneous work. The model used by the author for the aggregated forecast is an ARX a more causal model with lags of dependant variable using monthly data with dummy variables which control seasonality.

Chu (2009) presented international air transportation demand models in the form of a tourism approach. The forecast is made using several time-series (specifically ARMA) models to benefit of the long memory of the series. In his paper he developed the ARIMA and gave tips to remove seasonality without using dummy variables. Also criticizes about the accuracy of the MAPE and RMSE metrics and use it to compare the performance of the forecast in several periods.

Karlaftis (2008) presented a very complete work about forecast for both passenger traffic and flights in regional Greek airports. Again the issue of seasonality worked around a GARCH model to control that volatility disturbance. He then concludes that the "affects of the efficiency of parameter estimates and leads to biased – and erroneous – elasticity measures. *These errors could lead to significant underestimation of future demand at airports*".

Marazzo *et. al*. (2009), made a time-series analysis for air transportation demand and growth in Brazil. They use monthly data to forecast aggregated data in an Error Correction model, their emphasizing the interaction between the domestic passenger and the GDP their main purpose was to show that this relationship of the variables show a reactivation of the market after the crisis in order to provide better planning.

2.3. Structural Break Analysis

Traditionally the estimation of a regression model is based on the assumption that the means and variances of the error of these tested variables are constant over the time. Variables whose means and variances change over time are known as non-stationary or unit root variables. Therefore, incorporating non-stationary or unit root variables in estimating the regression equations gives misleading inferences.

Instead, if variables are non-stationary, the estimation of long-run relationship between those variables should be based on the co-integration method. Since the testing of the unit roots of a series is a precondition to the existence of co-integration relationship. Originally, the Augmented Dickey-Fuller (1979, 1981) test was widely used to test for stationarity. However, Perron (1989) demonstrated a failure in allowing for an existing structural break leading to a bias that reducing the ability to reject a false unit root null hypothesis.

Perron (1998) argues that most macroeconomic series are not characterized by a unit root but rather that persistence arises only from large and infrequent shocks, and that the economy returns to deterministic trend after small and frequent shocks.

This structural break can occur in time series data or cross sectional data, when there is a sudden change in the relationship being examined. A data can be found to be non-stationary if it has a unit root, or if it includes a structural break, before and after which data shows different patterns. Most tests that attempt to distinguish between a unit root and a (trend) stationary process will favor the unit root model when the true process is subject to structural changes, but is otherwise (trend) stationary within regimes specified by the break dates. Also, most tests trying to assess whether a structural change is present will reject the null hypothesis of no structural change when the process has a unit root component but also constant model parameters. Accordingly, there is voluminous literature on testing for a unit root under structural break(s).

Zivot and Andrews (1992) and Perron (1997) proposed determining the break point endogenously from the data, giving break dates as a by-product, but they are not as efficient as the break estimators. Later Christiano (1992) criticizes known date assumption as data mining arguing that the data based procedures are typically used to determine the most likely location of the break, i.e. by pre-test examination of the data, and this approach invalidates the distribution theory underlying conventional testing.

The Chow test shows whether the series has a break in the tested date. This first category of estimating breaks use the hypothesis of known date. Tests like Quandt-Andrews look for the presence of a break in the series, which may exist at any time within the sample period. The tests in the last category are in fact estimators, they first estimate the unknown date of the break, then test it.

Although to understand the basics of the structural break, it is better to start with the Chow Test. Instead we use estimators that are used to find unknown break dates and test them, because unknown date estimators that use more complicated tests basically rest on the same principles as the Chow test. Whether splitting data from the possible break point and estimating two generated sub-samples separately by least square gives significantly better than using the whole sample at once; if the answer is yes, the null hypothesis of no break is rejected.

3. METHODOLOGY

This section describes the evaluation process to select and validate the champion model for the dataset, describing the out-of-sample methodology and the metrics used to discriminate between the forecast accuracy of the time-series models. Also it present the variables used in the models, both the primary variables (endogenous and exogenous) and the secondary ones used to aid in the forecast of the exogenous variables, with the intention of present forecast to 2014 for the selected champion model.

3.1. Out of Sample Forecast

This procedure is used to discriminate between models the most accurate in the sample. The experimental process separates the sample in two areas the first is used to model the function and the second to forecast. The in-sample forecast is used to fit the model. In this stage all the econometric and statistical assumptions enter to obtain the most efficient model of the type (in our case time-series models of one and multiple equations and regimes). The out-of-sample forecast is used to convey the error in a metric to compare between models. For this the most used metrics are used:

The Mean Absolute Percentage Error (MAPE)

This metric obtain the measure of the absolute percentage error of each point, the result can be interpreted as a total error deviation which facilitates its interpretation in any unit, a weak point of this metric comes from the mathematical interpretation, the negative error only can be -100% but the positive can be infinite, thus misinterpreting the error, a strong point comes as the deviation show a standardized result.

$$MAPE = \frac{1}{K}\sum_{t}^{K}\left|\frac{\hat{X}_t - X_t}{X_t}\right| \times 100$$

The Root Mean Square Error (RMSE)

This metric solve the problem of the cancelled errors the same way the leas squares solved the curve optimization paradigm, the RMSE square root the measures the error square, reproduce a value with its units to compare between, the higher the error the bigger the penalization, that is why often this metrics present big variation, as the value comes with its units is often hard to explain the meaning of the variation, in this topic the MAPE is superior.

$$RMSE = \sqrt{\frac{1}{K}\sum_{t}^{K}\left(\hat{X}_t - X_t\right)^2}$$

3.2. Variables and Forecast Assumptions

The endogenous variable used to model demand is the Revenue Passenger per Kilometer (RPK). The modeling doesn't care about the type of passenger as this work is solely to study aggregated demand. Hence a unit that doesn't care about types is the RPK, the final result will be stripped of the kilometer component by dividing for the average stage flight, only to provide a more common unit, it will also be useful to the metric of discrimination.

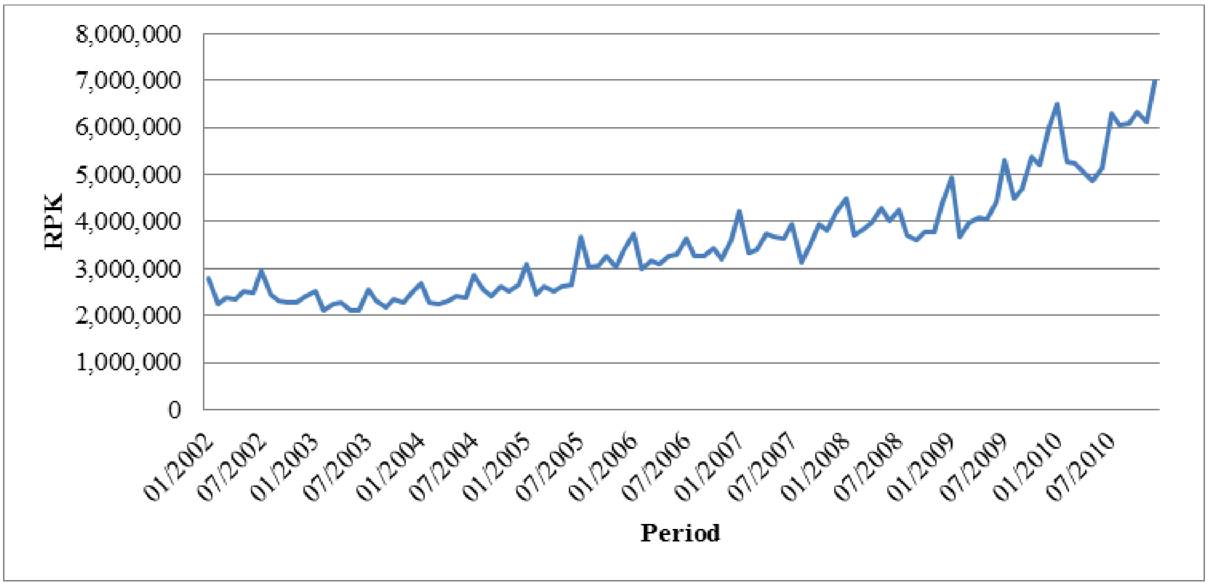

Figure 1 – Endogenous Variable RPK

When needed the variable would be transformed to logarithms and/or integrated to provide better fit on the models. In Figure 1 can be observed that the variable don't have stationary behavior which is comprehensible as it embodies many shocks with the increasing demand of the last four years, usually the presence of unit root is lifted by first difference but the Augmented Dickey-Fuller (ADF) test presented in Table 1 shows the contrary:

Table 1 - Augmented Dickey-Fuller test statistic for DRPK

Null Hypothesis: D(RPK) has a unit root
Exogenous: Constant
Lag Length: 11 (Automatic based on SIC, MAXLAG=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.178639	**0.2154**
Test critical values:	1% level	-3.500669	
	5% level	-2.892200	
	10% level	-2.583192	

*MacKinnon (1996) one-sided p-values.

The Augmented Dickey-Fuller test in second differences presented in Table 2, reject the null hypothesis of unit root, so in order to utilize stationary series when needed, the second order integration will be need.

Table 2 - Augmented Dickey-Fuller test statistic for D²RPK
Null Hypothesis: D(RPK,2) has a unit root
Exogenous: Constant
Lag Length: 10 (Automatic based on SIC, MAXLAG=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-8.033483	**0.0000**
Test critical values:	1% level	-3.500669	
	5% level	-2.892200	
	10% level	-2.583192	

*MacKinnon (1996) one-sided p-values.

The persistency of non-stationarity behavior can also be explained with seasonal variables and else a different degree of integration between the first and second level, which allow us to study more complex time-series models.

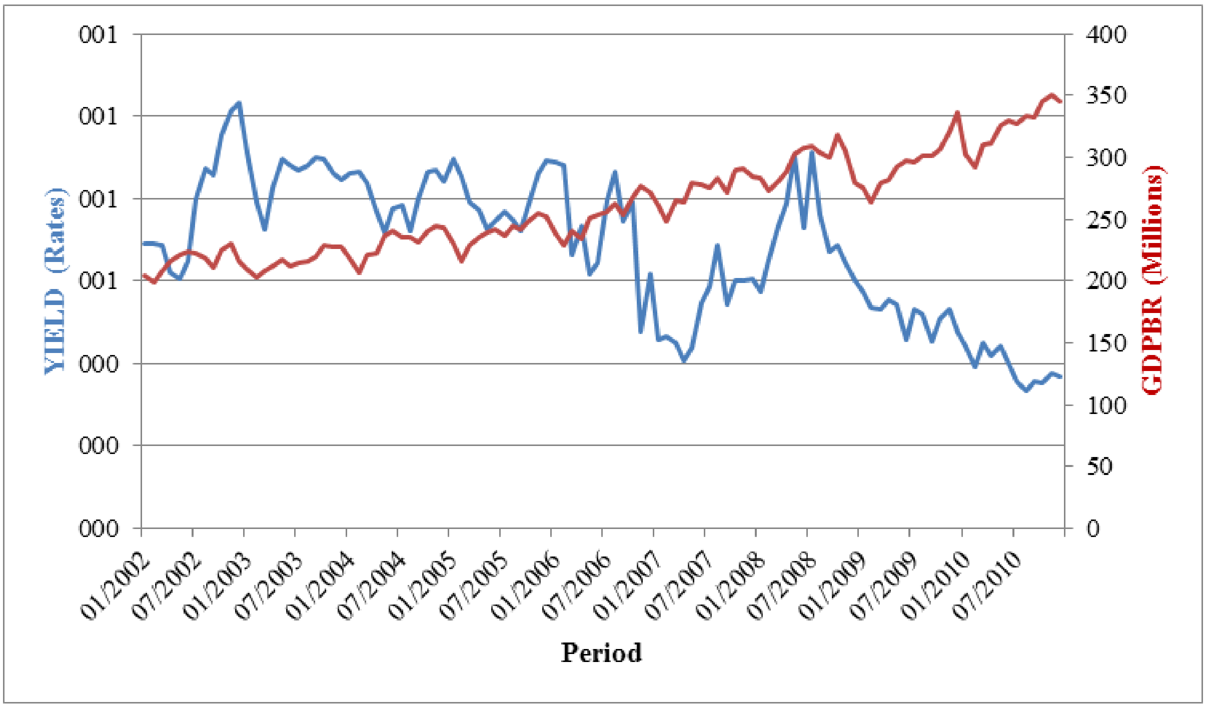

Figure 2 – Exogenous variables GDPBR, YIELD

The exogenous variables are the Gross Domestic Production of Brazil (GDPBR) and the revenue per passenger transported (YIELD), in a simpler they represent the aggregated income and the aggregated price level respectively, which make them perfect demand variables. As seen on Figure 2 the GDPBR is in millions and the yield in rates.

They are also not stationary for their behavior contains several exogenous shock, we test for unit root in first differences with several methods similar to the ADF:

Table 3 – Unit Root test statistic for exogenous variables

Group unit root test: Summary
Series: GDPBR, YIELD
Sample: 2002M01 2010M12
Exogenous variables: Individual effects
Automatic selection of maximum lags

Null: Unit root (assumes individual unit root process)

Im, Pesaran and Shin W-stat	-13.2134	**0.0000**	2	203
ADF – Fisher Chi-square	108.792	**0.0000**	2	203
PP - Fisher Chi-square	156.567	**0.0000**	2	212

So we reject the unit root hypothesis of non-stationary behavior in first differences, this should be useful when modeling time-series with exogenous variables.

4. CASE STUDY: BRAZILIAN AIR TRANSPORTATION DEMAND

4.1. One Regime Models

M1 – Causal Model

The model takes the form of a Cobb-Douglass log-linear transformation, the parameter estimated are fixed elasticities, which gives the RPK an economy of scale characteristic. This simple model although with efficient parameters and great goodness of fit has a fatal flaw, the non-stationary series, which turn the model inconsistent and biased. Nevertheless let's observe the behavior of the forecast on the metric analysis:

Now we use this non stationary model of Table 4 to analyze the presence of structural break as an alternative of the unit root analysis, the test necessary to reject that hypothesis is only doable while in OLS.

Studying the Figure 4 we notice a jump in the error term around the second half of 2005, this is explained by the Re-regulation that began in 2003, as the Brazilian aviation authorities implement new measures to manage the steady growth of the market, this re-regulation defined controlling an alleged excess capacity route and over-competition in the market but the authorities are to impose restrictions only if carriers are in a poor financial situation.

Table 4 – OLS estimated parameters

Dependent Variable: LRPK
Method: Least Squares
Sample: 2002M01 2008M12
Included observations: 84

Variable	Coefficient	t-Statistic	Prob.
C	6.789209	11.68	**0.0000**
LYIELD	-0.169852	-2.65	**0.0096**
LGDPBR	1.464971	13.67	**0.0000**
R-squared	0.773494	Adjusted R-squared	0.767901
Akaike info criterion	-1.622702	Durbin-Watson stat	1.283771

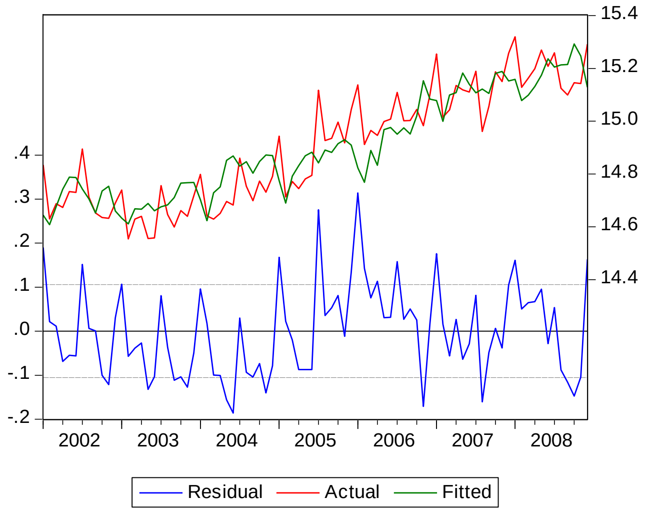

Figure 3 – Actual and Fitted of Causal Model, and Residuals

To assure theoretically this "jump" we ran a Quandt-Andrews unknown breakpoint test, a test that find breakpoint through the series point by point, this test uses the same methodology of the Chow test but comparing recursive parameters change:

Table 5 – Quandt-Andrews unknown breakpoint test

Null Hypothesis:	No breakpoints within data	
Varying regressors:	All equation variables	
Equation Sample:	2002M01	2010M12
Test Sample:	2003M06	2009M07
Number of breaks compared:		74
Statistic	Value	Prob.
Maximum LR F-statistic (**2005M07**)	7.038881	0.5184
Maximum Wald F-statistic (**2005M07**)	7.038881	0.5184
Exp LR F-statistic	1.750226	0.4761
Exp Wald F-statistic	1.750226	0.4761
Note: probabilities calculated using Hansen's (1997) method		

The test tells us to reject the hypothesis of no breakpoint on the month 7 of 2005, just around the date we suspected, as we will see on the following time-series models, around that date the error term should "jump" showing presence of structural break even on apparently stationary series.

M2 – ARMAX(2,2) – ARMA(1,1)

The model is an Autoregressive and moving average model that lacks of the first lagged AR and MA variables, this was result of an inside test of many ARMA models, in order to choose one that pass the tests of individual and global significance, also be noted that the AR(2) component is not coincidence, it may be explained as an auxiliary parameter to give stability, the representation is:

Table 6 – ARMAX estimated parameters

Dependent Variable: LRPK Method: Least Squares Sample (adjusted): 2002M03 2008M12 Included observations: 82 after adjustments			
Variable	Coefficient	t-Statistic	Prob.
C	6.973002	9.83	**0.0000**
YIELD	-0.232746	-2.25	**0.0271**
LGDPBR	1.47237	12.07	**0.0000**
AR(2)	-0.599955	-3.50	**0.0008**
MA(2)	0.848842	7.60	**0.0000**
R-squared	0.794965	Adjusted R-squared	0.784314
Akaike info criterion	-1.667617	Durbin-Watson stat	1.411794

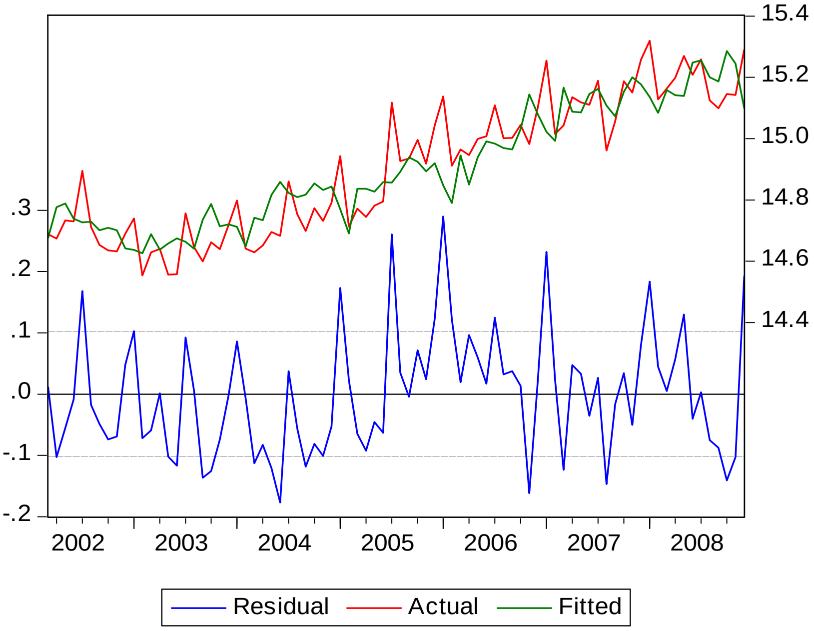

Figure 4 – Actual and Fitted of ARMAX Model, and Residuals

Figure 5 depict a better fit to the actual values of the model, but the presence of structural break is showing as a jump in the error term is clear around 2005, but this deviation is lower than in the linear model as the AR(2) and MA(2) components help smooth this trade-off. The out of sample forecast curve of the ARMAX presented in Figure 6 is nearer to the actual values, and as we will see the desirable is to be almost over that, as there is a tendency of underestimation because the exogenous and historical lags can't explain the shocks at the same rate as they occur, look at the pike around 2010, the model reach that behavior but to early in comparison to the actual.

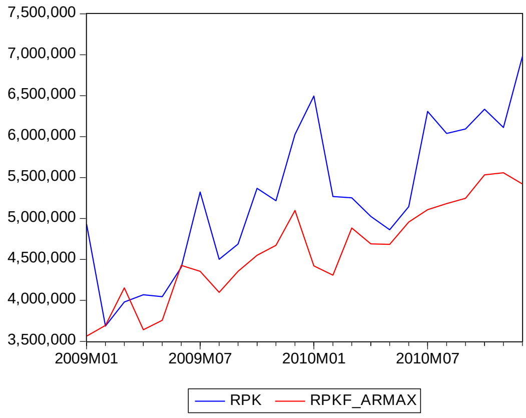

Figure 5 – Actual and Forecast of ARMAX Model

M3 – ARIMAX(2,2,1)

Early in the study of the endogenous variables was discussed that a stationary process of the RPK can be obtained with second order differences, again using logarithms to smooth the relation between the exogenous variables, after a Box-Jenkins filter we came up with an AR(2) and MA(1) components which better fitted the series.

Table 7 – ARIMAX estimated parameters

Dependent Variable: D2LRPK Method: Least Squares Sample (adjusted): 2002M05 2008M12 Included observations: 80 after adjustments			
Variable	Coefficient	t-Statistic	Prob.
C	-0.010315	-2.15	0.0349
YIELD	0.010889	1.93	0.0571
DLGDPBR	0.476419	2.56	0.0124
AR(1)	-0.579387	-5.27	0.0000
AR(2)	-0.403662	-3.66	0.0005
MA(1)	-0.969647	-68.42	0.0000
R-squared	0.739961	Adjusted R-squared	0.722391
Akaike info criterion	-1.836881	Durbin-Watson stat	1.932789

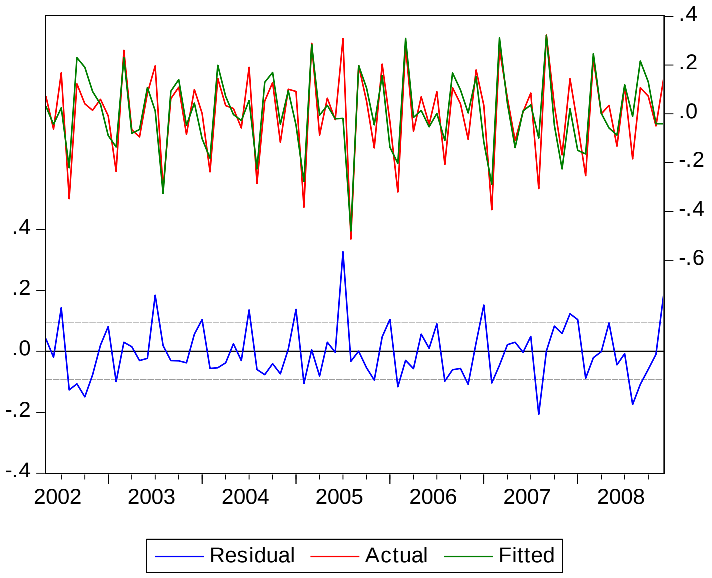

Figure 6 – Actual and Fitted of ARIMAX Model, and Residuals

From Figure 7 is possible to se that the fitted curve is closer to the actual in the in-sample forecast, and the super-position on the out-of-sample forecast.

Also Figure 8 shows the jump on the 2005 errors which are lower as the residual component is more stable, but the forecast show an exaggerated behavior around the actual, almost destabilized.

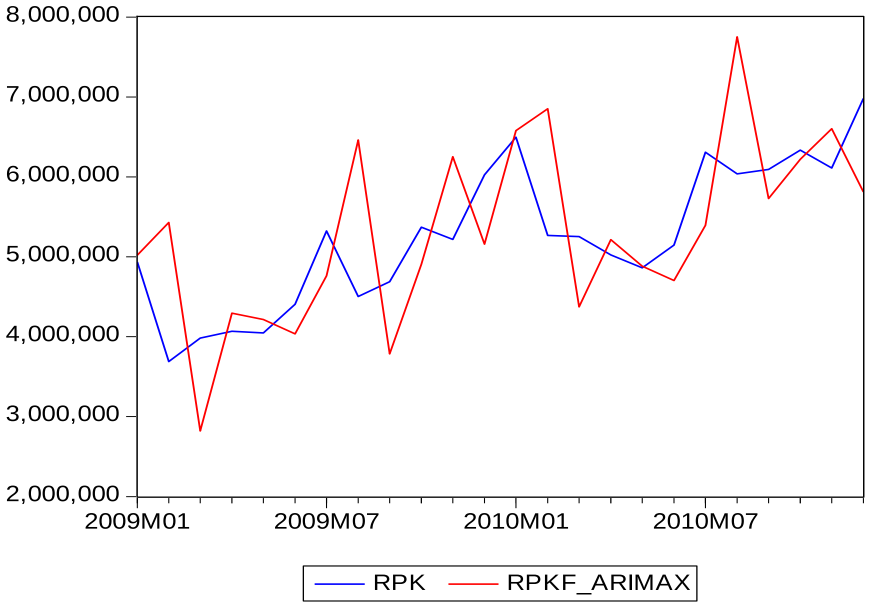

Figure 7 – Actual and Forecast of ARIMAX Model

M4 – SARIMAX(0,1,2) (2,1,0) $_6$

A SARIMAX model is a generalization which includes a seasonal component. After observing the behavior of the correlogram (Figure 9) of the first differences of the LRPK series we choose a 6th difference for the mix autoregressive component.

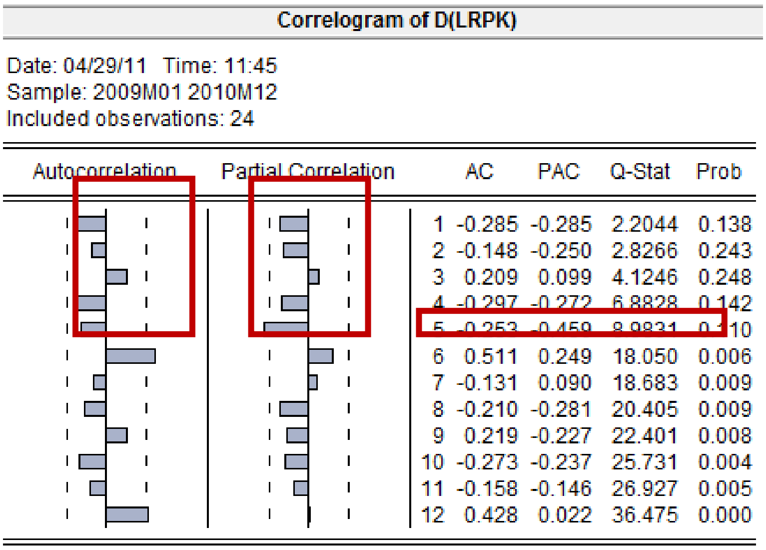

Figure 8 – Correlogram of LRPK first difference

The range of the AR and MA terms were chosen the same way of the ARMA, after a long test of individual and global significance Box-Jenkins filter, the representation with the better parameter consistency were:

Table 8 – SARIMAX estimated parameters

Dependent Variable: D6LRPK Method: Least Squares Sample (adjusted): 2002M09 2008M12 Included observations: 76 after adjustments			
Variable	Coefficient	t-Statistic	Prob.
C	0.095185	2.09	0.0401
YIELD	-0.078253	-1.31	0.1930
DLGDPBR	-0.063311	-0.36	0.7191
AR(1)	-0.461045	-3.92	0.0002
AR(2)	-0.273345	-2.36	0.0209
MA(1)	0.957979	21.50	0.0000
MA(2)	0.973797	34.87	0.0000
R-squared	0.380764	Akaike info criterion	- 2.316092
Adjusted R-squared	0.326917	Durbin-Watson stat	1.863239

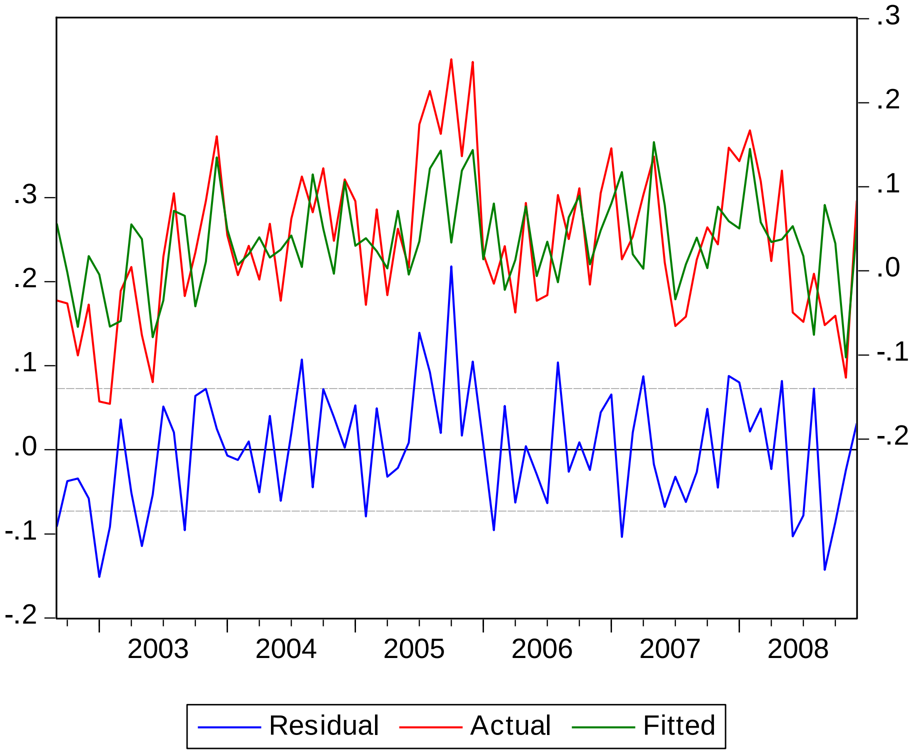

Figure 9 – Actual and Fitted of SARIMAX Model, and Residuals

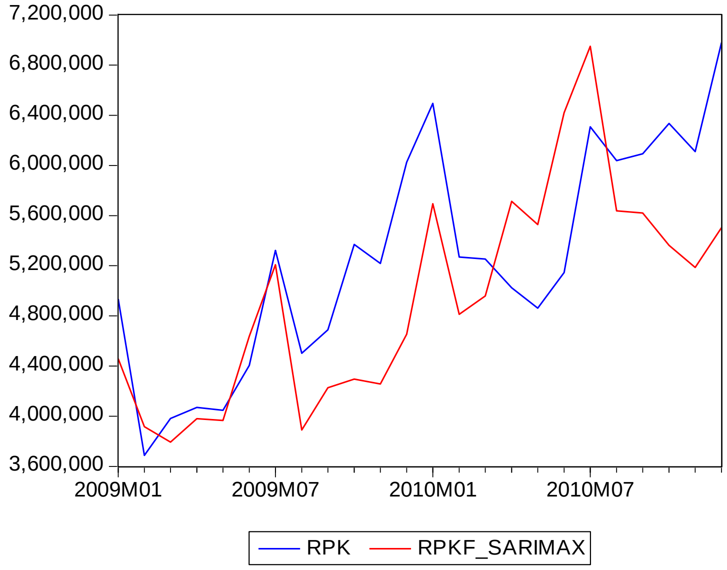

Figure 10 – Actual and Forecast of SARIMAX Model

In Figure 10, the in-sample, the fitted model follows the actual model very close, except one again for the period at beginning of 2005, while the Figure 11 show that the seasonal behavior is very similar but unable to accompany the full effect.

M5 – ARFIMAX(2,d,0)

Remembering that the endogenous variable reach stationarity at second differences, maybe it will in some point between, so we test a fractionary ARIMAX on first differences, the non-linear estimation begins with a smoothing of the d-parameter, obtaining a [0<d<1] value between zero and one, proving our suspects:

Table 9 –ARFIMA(2,d,0) estimated parameters

Dependent Variable: DLRPK Method: Non Linear Least Squares Sample (adjusted): 2002M03 2008M12 Included observations: 82 after adjustments			
Variable	Coefficient	t-value	p-value
d parameter	0.32903	1.960	0.053
AR-1	- 0.65521	- 3.340	0.001
AR-2	- 0.28560	- 1.820	0.073
Constant	0.00329	1.900	0.061
DLYIELD(t)	0.03161	0.402	0.689
DLYIELD(t-1)	- 0.15229	- 1.930	0.058
DLGDPBR(t)	- 0.46175	- 1.740	0.087
DLGDPBR(t-1)	1.24527	4.390	0.000

Although the lag of the first difference of the YIELD variable were not significant, this doesn't undermine our efforts, the AIC (Akaike Info Criterion) is a criteria to compare between similar models, the sigma is the standard deviation which is low compared to the standard error of the variables.

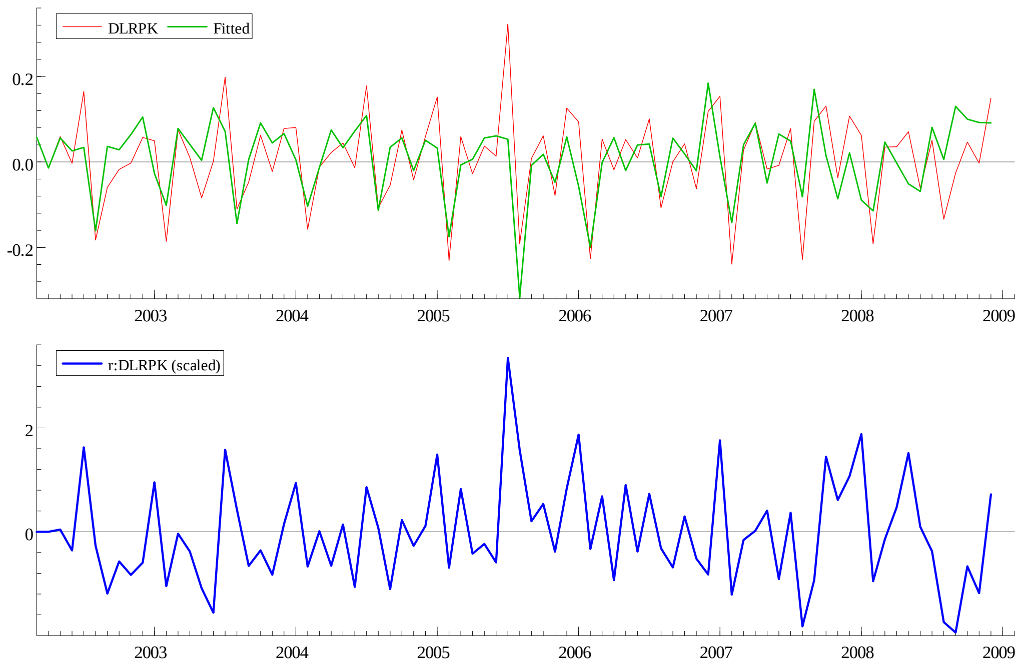

Figure 11 – Actual and Fitted of ARFIMAX Model, and Residuals

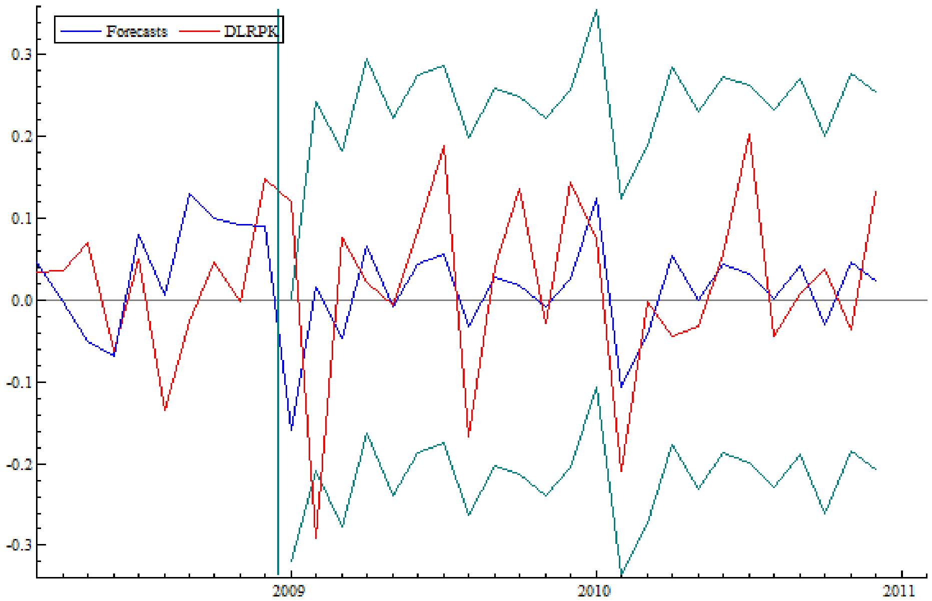

Figure 12 – Actual and Forecast of ARFIMAX Model

In figure 12 the presence of an error jump is still visible around 2005. The actual and fitted of the in-sample, and the forecast curve of the out-of-sample (Figure 13) indicate that this model apparently is one of the most graphically efficient.

M6 – VAR

The VAR need the series to be stationary so, we use as endogenous the second difference of the LRPK, and the first difference of the others as we saw the unit root test earlier, here the modeling take two steps, the base model have two equations and be called VAR2, the second model have three equations and be called VAR3, the estimation of the models give structural efficiency and strong consistent parameters as shown on Tables 10 and 11:

Table 10 – VAR2 estimated parameters

Sample (adjusted): 2002M05 2008M12 Included observations: 80 after adjustments		
	Dependent Variable	
Exogenous Variable	D2LRPK	DLYIELD
D2LRPK(-1)	-0.899779*	0.083051°
D2LRPK(-2)	-0.552748*	0.00717
DLYIELD(-1)	-0.110705	-0.198097*
DLYIELD(-2)	-0.207467*	0.183076*
C	0.002259	-0.001408
DLGDPBR	0.225542°	0.273924°
R-squared	0.563767	0.104646
Adj. R-squared	0.534291	0.044149
Akaike AIC	-1.319533	-1.316309
(°) 10% Significance, (*) 5% Significance		

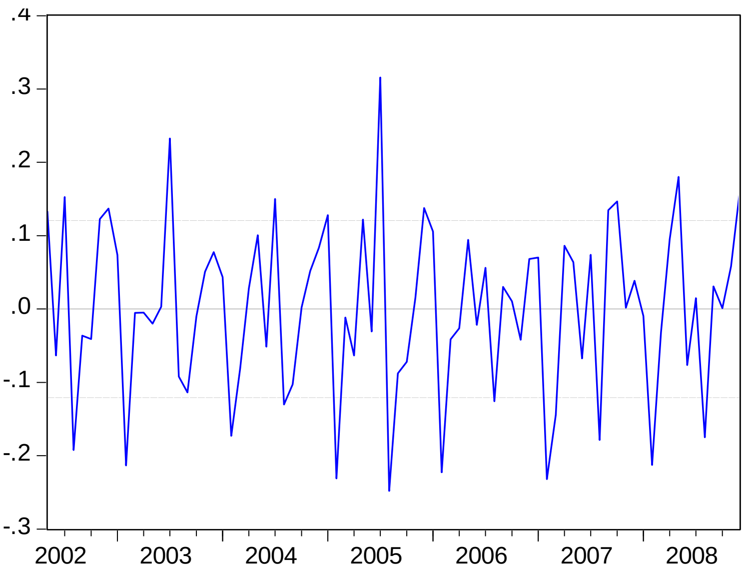

Figure 13 –VAR2 Model Residuals

Figure 14 and 15 explain that residues escapes the control bands irregularly; the model should have global consistency, the VAR3 control better this irregularity, but the over-parameterization is not recommendable:

Table 11 – VAR3 estimated parameters

Vector Autoregression Estimates Sample (adjusted): 2002M05 2008M12 Included observations: 80 after adjustments			
	Dependent Variable		
Exogenous Variable	D2LRPK	DLYIELD	DLGDPBR
D2LRPK(-1)	-1.028891*	0.021074°	-0.025459*
D2LRPK(-2)	-0.692634*	-0.068466	-0.040272*
DLYIELD(-1)	-0.142825*	-0.217824*	-0.021018°
DLYIELD(-2)	-0.236099*	0.156416*	-0.058568*
DLGDPBR(-1)	1.096913*	0.364929°	-0.021036
DLGDPBR(-2)	1.562228*	0.847557*	-0.070026
C	-0.01104	-0.007029	0.003838
R-squared	0.682694	0.157587	0.072615
Adj. R-squared	0.656615	0.088347	-0.003608
Akaike AIC	-1.612846	-1.352257	-3.695926
(°) 10% Significance, (*) 5% Significance			

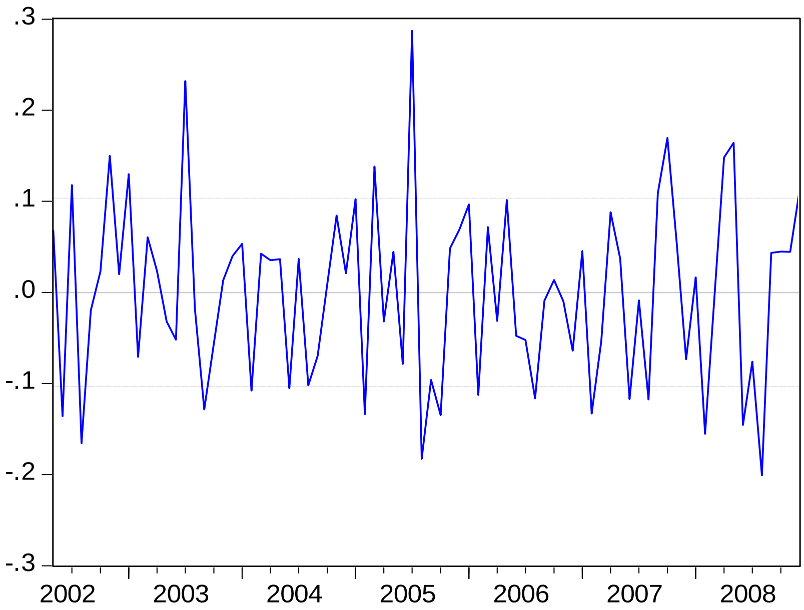

Figure 14 – VAR3 Model, and Residuals

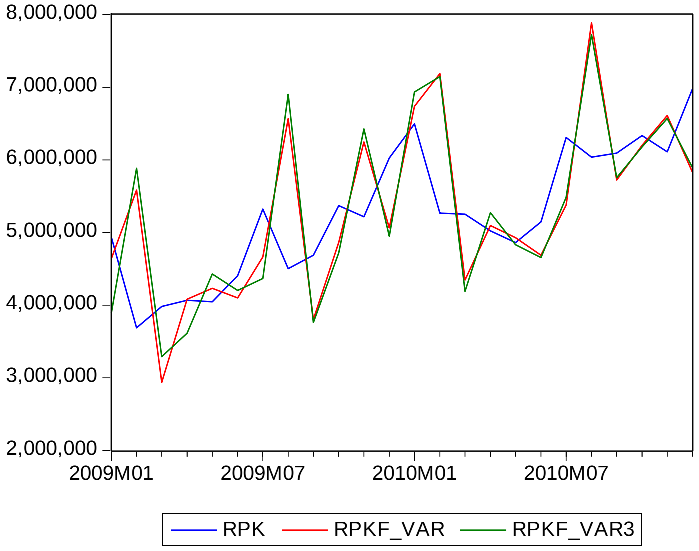

Figure 15 – Actual and Forecast of VAR and VAR3 Model

Figure 16 shows that the VAR3, controlling better the residuals out-perform the base VAR2 model, the forecast line are almost overlapping, while is better to choose less variables to avoid problems of over-parameterization, the benefits of doing so are considerable.

M7 – VEC

We chose to measure the relation between the RPK and YIELD, using the GDPBR as an exogenous variable, after following the methodology described in the paper of Marazzo et al. (2009) who depict a relation between the RPK and GDP, after replicating that model and also test a three equation VEC which RPK, GDP and YIELD, the results show that the initial model have better information criteria and the behavior of the series adjust better with the out-of-sample data, the same as the VAR.

Table 12 – VEC estimated parameters

Vector Error Correction Estimates Sample (adjusted): 2002M04 2008M12 Included observations: 81 after adjustments		
Cointegrating Eq:		
LRPK(-1)	1.00000	
LYIELD(-1)	0.54823*	
C	-14.7325	
Error Correction:	D(LRPK)	D(LYIELD)
CointEq1	-0.272118*	-0.345301*
D(LRPK(-1))	-0.355107*	0.332187*
D(LRPK(-2))	-0.237557*	0.127759
D(LYIELD(-1))	0.054178	-0.087677°
D(LYIELD(-2))	-0.068504	0.273525*
C	-2.029064*	-2.273623*
LGDPBR	0.370337*	0.412027*
R-squared	0.360931	0.191747
Adj. R-squared	0.309115	0.126212
Akaike AIC	-1.924455	-1.400361
(°) 10% Significance, (*) 5% Significance		

We use logarithms to aid smooth the relation between variables, in between {} is the cointegration vector, which attains stationarity as shown in Table 12, allowing us to study directly the relation of the variables.

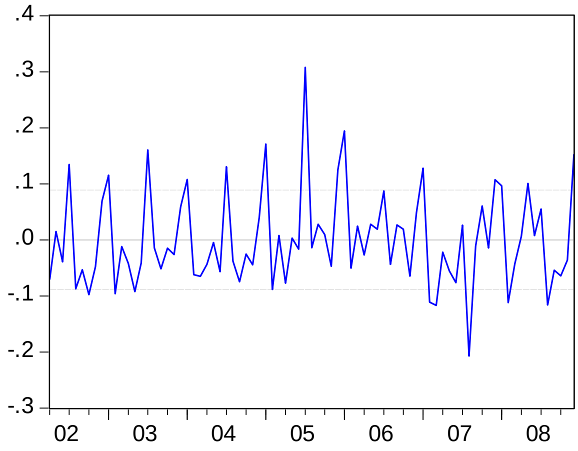

Figure 16 – VEC Model Residuals

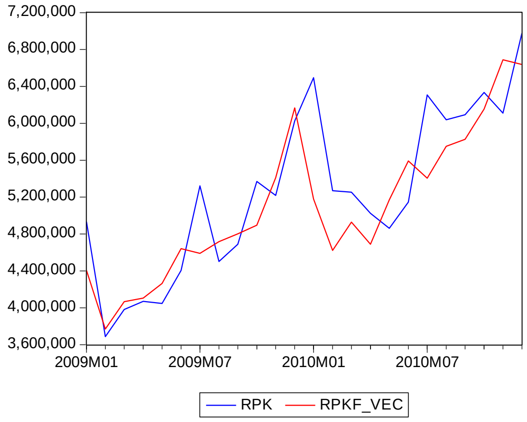

Figure 17 – Actual and Forecast of VEC Model

Figure 17 show the residual better controlled and as seen in the out-of-sample forecast vs. actual of Figure 18, the fit depict a good curve behavior:

4.2.Multiple regime models

M8 – STAR

To aid us in the modeling was used the JMulti software which began testing nonlinearity and selecting the type of transition function, a Logit of order 1 (LSTR1) the endogenous includes an AR(1) component and the exogenous variables includes first lagged components.

Table 13 – STAR estimated parameters

Variables in AR part:	C DRPK(t-1) DGDPBR(t) DYIELD(t) DGDPBR(t-1) DYIELD(t-1)		
Transition variable: Sample range: Transition function:	DGDPBR(t) [2002M3,2008M12], T=82 LSTR1		
Variable	Estimate	t-stat	p-value
----- linear part ------			
CONST	247,249.4	0.0019	0.0019
DRPK(t-1)	-0.3868	-1.3871	0.0788
DGDPBR(t)	-55,857.8	-3.3374	0.0012
DYIELD(t)	1,203,730.7	0.6917	0.4909
DGDPBR(t-1)	13,084.6	1.5186	0.1323
DYIELD(t-1)	4,692,088.6	2.5211	0.0134
---- non linear part ----			
CONST	930,656.9	3.6626	0.0004
DRPK(t-1)	0.1759	0.5915	0.0556
DGDPBR(t)	45,538.0	2.5564	0.0122
DYIELD(t)	-935,155.9	-0.5113	0.6104
DGDPBR(t-1)	-14,117.0	-1.4255	0.1574
DYIELD(t-1)	-5,823,185.0	-2.9518	0.0040
Gamma	15.3	1.9780	0.0552
C1	-7.3	-8.4561	0.0000
R2:	3.751E-01		
adjusted R2:	3.692E-01		

There were three choices for the transition variables, we choose $GDPBR_t$ because it had the better parameter efficiency although as can be seen on Table 13 the parameter of the YIELD was significant only on the lagged part, this doesn't rest power to the model, the importance here is that the transition function had good parameters and the model rejects persistency of non-linearity. In this case we didn't use logarithm transformation to improve the transition between exogenous variables, which explains why the parameter of the Yield is in the millions and in part its non-significance.

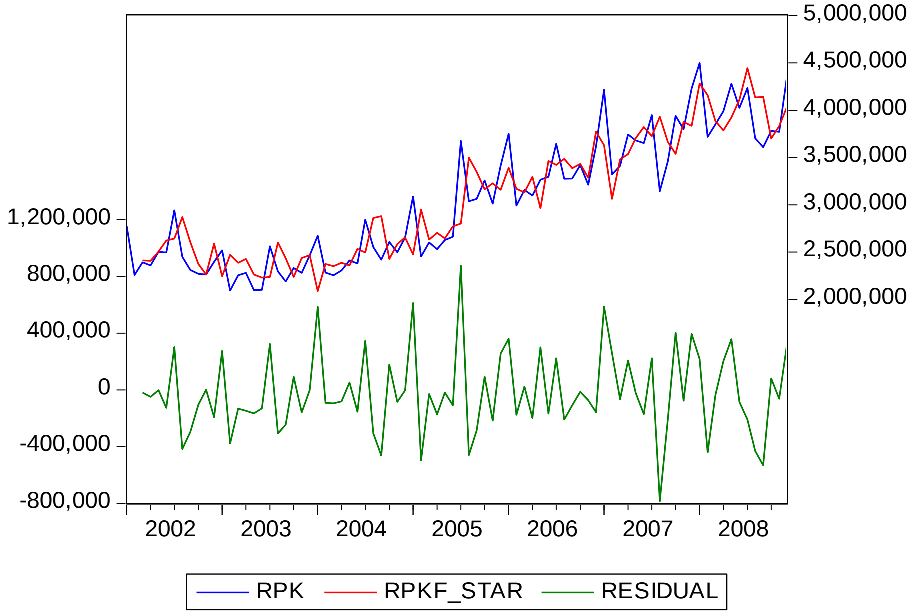

Figure 18 – Actual and Fitted of STAR Model, and Residuals

To understand Table 13 we need to remember the representation of the STAR models observed on the literature review, there was a part of the equation for linear components and other for non-linear, both multiplied by a homogenization of the transition variable, assuring a threshold between stages which creates the Self-Exciting models, the information criteria is used to compare between STAR, that was the primary tool to choose this estimation over the other possibilities created by further transition variables.

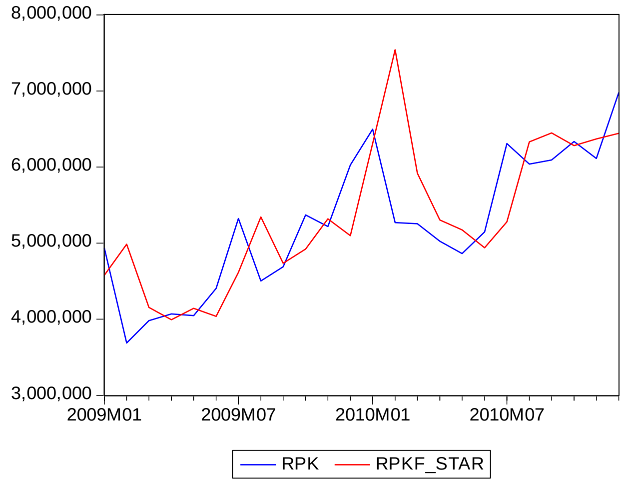

Figure 19 – Actual and Forecast of STAR Model

The LSTAR model obtained seems to perform better than the one-regime models.

M9 – MSAR

The Markov-switching model studied is a MSAR(2,2,0) a two-regime ARMA(2,0), (the AR component was obtained trough the same Box-Jenkins filter of the ARIMA models) the model can detect the regime and switch the intercept after a transition probability, the same probability is used to detect the better scenario of the forecast regime, there is a wide variety of Markov models, there are those who switch ARMA components, and even variance, for this model was chosen a somewhat simple MSAR because it explain very well the behavior of the RPK as it detect the structural break of 2005m7 and switch to another regime. This repeats again prior to 2008 and again on 2009, dates when the bigger increase on demand occurred:

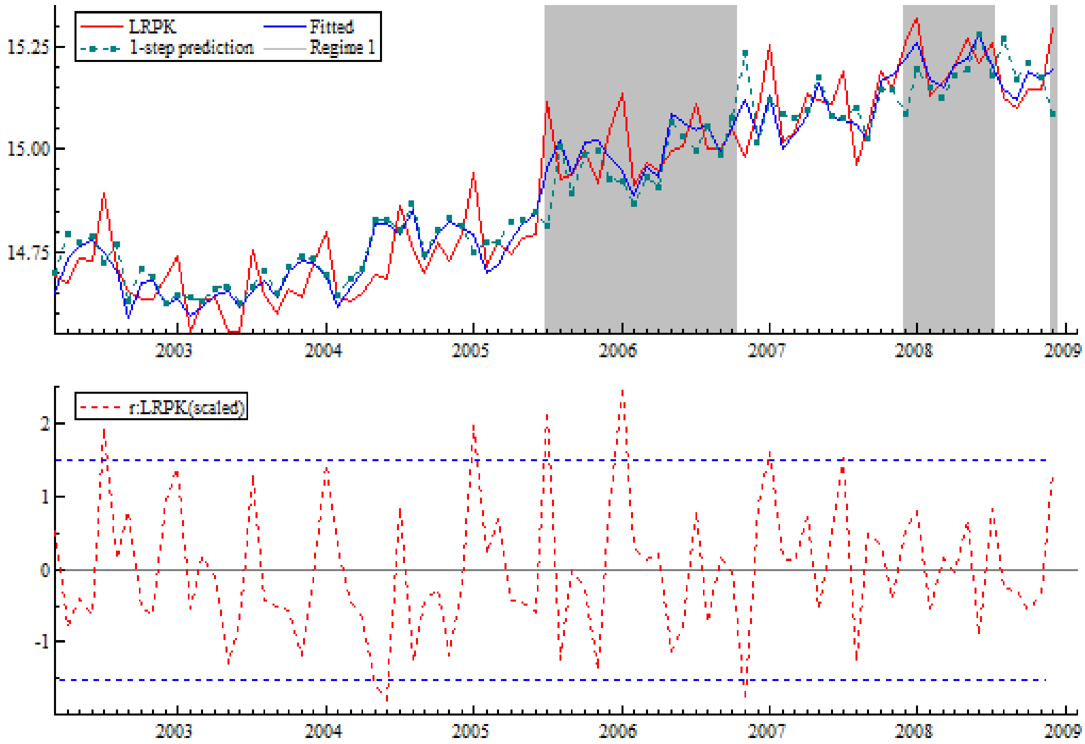

Figure 20 – Actual and Fitted of MSAR Model, and Residuals

Figure 21 presents residuals better controlled, the estimated parameters are on logarithms and as we can see in the Table 14 almost all the parameters attain a significance of less than 1%.

Table 14 – MSAR estimated parameters

Switching(1)	Modelling LRPK by MS_ARMA(2,2,0) estimation sample is: 2002(1) - 2008(12)				
	Coefficient	Std.Error	robust-SE	t-value	t-prob
AR-1	0.135916	0.1255	0.1072	1.270	0.209
AR-2	-0.250891	0.1197	0.1099	-2.280	0.025
LYIELD	-0.248182	0.06296	0.07017	-3.540	0.001
LGDPBR	1.25261	0.1153	0.1046	12.000	0.000
Constant(0)	7.88348	0.6181	0.5538	14.200	0.000
Constant(1)	8.03679	0.6312	0.5731	14.000	0.000
sigma	0.0760697	0.008986	0.009979	7.620	0.000
p_{0|0}	0.929984	0.05951	0.08241	11.300	0.000
p_{0|1}	0.131201	0.1007	0.1241	1.060	0.294
log-likelihood		79.1597106	AIC		-1.71121245

The transition probabilities, p_{i|j} = P(Regime i at t+1|Regime j at t) are:

	Regime 0,t	Regime 1,t
Regime 0, t+1	0.92998	0.1312
Regime 1, t+1	0.070016	0.8688

And graphically:

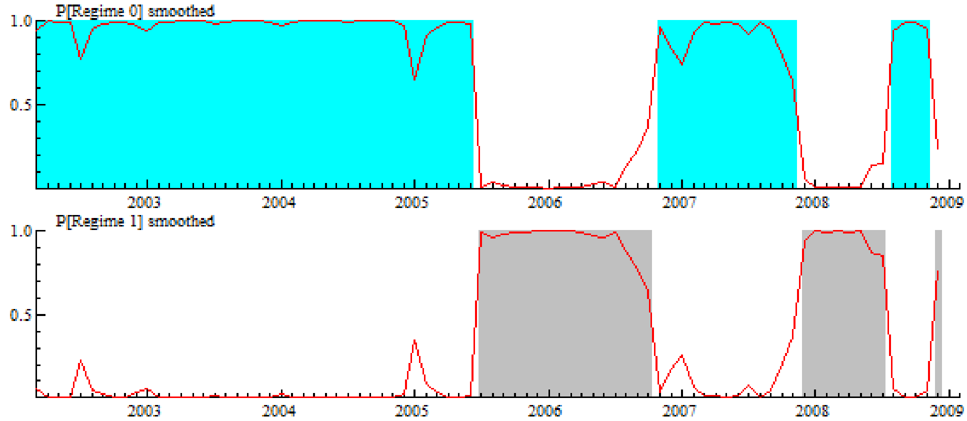

Figure 21 – Smoothed probabilities of MSAR regimes

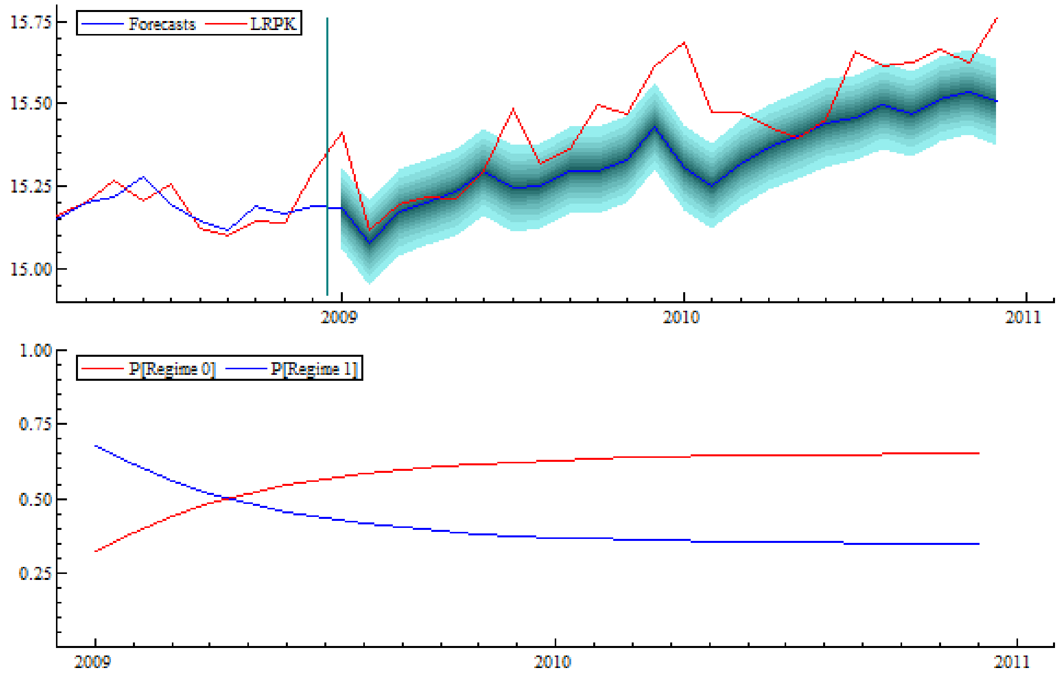

Figure 22 – Actual, Forecast and Probabilities of MSAR Model

Although the model better explains the behavior of the variable and its economic impact, the forecast appear not as accurate as the STAR model, or even other single-regime models, the probability of being on the second regime –the regime of increasing demand– which empirically explains better the stage, was rejected, the model predicts the first scenario as the most probable, which can be explained by the decreasing increase rate of the demand, but still.

Table 15 - Regime classification based on smoothed probabilities

	Sample	Months	Average probabilities
Regime 0	2002(03) - 2005(06)	40	0.974
	2006(11) - 2007(11)	13	0.902
	2008(08) - 2008(11)	4	0.971
	Total: 57 months - 0.6951%		
Regime 1	2005(07) - 2006(10)	16	0.945
	2007(12) - 2008(07)	8	0.956
	2008(12) - 2008(12)	1	0.762
	Total: 25 months - 0.3049%		

4.3.Comparison and Forecast

After the out-of-sample forecast, when the dependant variables are transformed we revert them to its original, removing logarithms and/or re-integrating the series, in order to produce the metrics described on chapter three, that will serve us to discriminate the accuracy of the models. Note that this selection does not mean that this model will always be the most effective, just for the selected sample within the database, for that we need to validate our results, to do so we need to stress the models. We achieve this redoing the in-sample modeling for another sample, say 2003 to 2009, for all the models and comparing the out-of-sample of 2010 recollecting another set of metrics. Hopefully the results of both experiments will be close, thus validating our selection of a “champion model”. Again it is not the intention of the paper to create a law, but a methodology to select this champion model.

We stated before that for the RMSE metric we were to rate the result by the average stage to withdraw the kilometer of RPK leaving an indicator of passengers, meaning that the value of the metric show the deviation in terms of passengers (thousands) per year, for the MAPE the value indicates the average percentage deviation from the mean per year.

Figure 23 – Out-of-sample Selection Metrics

	Causal	armax	arimax	sarimax	arfimax	var2	var3	vec	str	ms
RMSE	17.59	17.40	19.16	15.32	12.06	20.40	22.13	9.79	14.46	17.31
MAPE	11.98%	12.01%	14.52%	11.33%	9.06%	15.21%	17.37%	6.81%	9.68%	11.43%

The most accurate of the one-regime model is the VEC followed close by the ARFIMAX model. For the multiple regime models, we found the comment from the paper of Teräsvirta (2004) in which he says that STR models generally outperform MS intercept models like ours, either way the champion model should be the VEC and for the second best we choose the ARFIMAX as it performs better than the multiple regime models. In the validation we found similar results, while the difference between accuracy is smoothed for the broader sample; the estimated parameters were in many cases not as significant as in the previous modeling, but at least give us a sense of accomplishment observe a validation of the modeling stress.

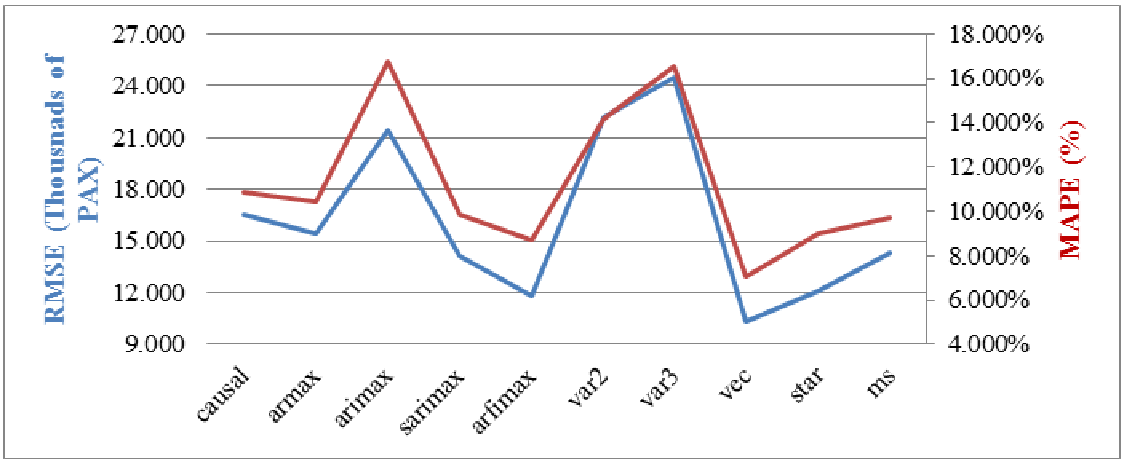

Figure 24 – Out-of-sample Validation Metrics

	Causal	armax	arimax	sarimax	arfimax	var2	var3	vec	str	ms
RMSE	17.59	17.40	19.16	15.32	12.06	20.40	22.13	9.79	14.46	17.31
MAPE	11.98%	12.01%	14.52%	11.33%	9.06%	15.21%	17.37%	6.81%	9.68%	11.43%

With these results we proceed to forecast RPK for the champion model, the VEC and also for the ARFIMAX as the second best and also to observe the behavior of the regime change analysis we analyze the behavior of the MSAR, and with the results give an estimated grow rate for the period 2011 to 2014.

The VEC model presents a seemingly steady forecast so the forecast have a steady growth of 12.5% per year, however the ARFIMAX model shows a lower growth rate of 10.16%, and the MSAR an even lower rate of 1.91%.

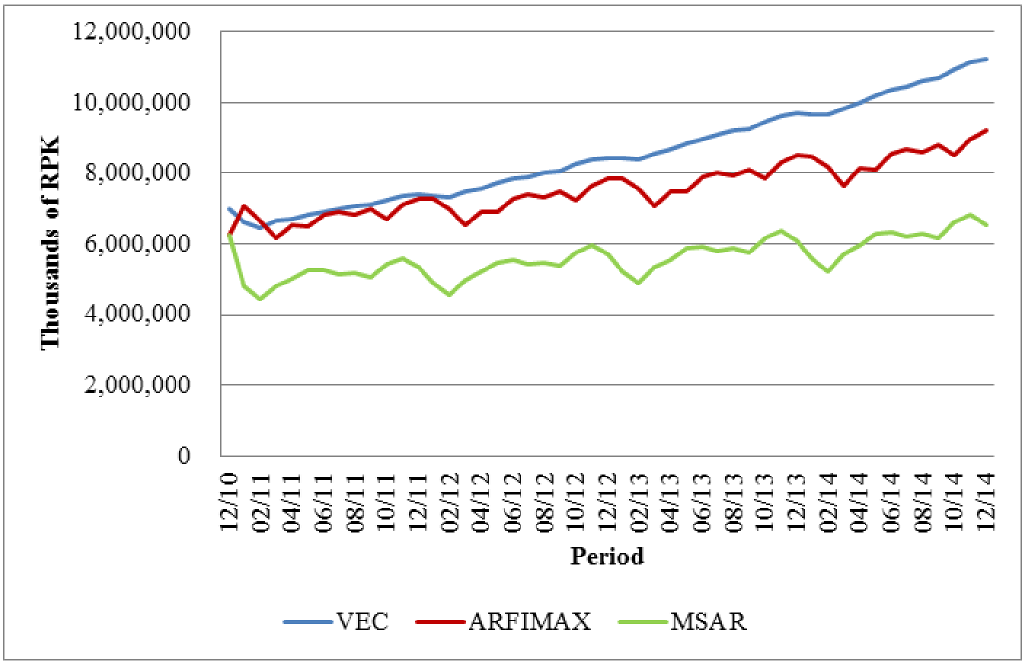

Figure 25 – Forecast to 2014, VEC and STAR

The reason for the MSAR decrease first is that the models put the forecast on the first regime, a regime of slower growth than the second; otherwise as we can see the tendency is nearly equal.

5. CONCLUSIONS

Brake date search tests should be used to pinpoint hidden structural breaks, because the unit root test/integration cannot remove entirely the nonstationary behavior of a series. The VEC methodology has heavy presence on the literature given its easy modeling and the granger causality analysis, specifically for the Brazilian analysis we have the contribution of Marazzo (2009) and Fernandez (2010) who study the relationship between the same variables used in this paper to analyze demand. In this way, we validate their efforts, given that our champion model is a Vector of Error Correction.

The metrics used in the selection and validation have been intensely used in the economic field (as in the biology and physics) and provide good indicators of accuracy, often the RMSE is given more importance because it penalize more the bigger deviations, but ultimately both metrics provide similar discrimination.

The forecast of the champion model give us a growth rate of roughly 8% and the second best of nearly 7%, given that in 2010 the World Bank show a economy grow rate in the Brazilian Real GDP of 7.5%, the models don't stray off the apparent way.

The weak point of the models comes around of the accuracy of the variables, the modeling begins with a hypothesis of precise exogenous data to achieve precise forecasts, and it should be interesting redo the work with final results when available.

When studying demand it's recommended to analyze the stationarity of the series' first because robust forecast can't be done without this, an alternative is the error correction vector which allow us to use of a stationary linear combination of the series, also using a Markov-switching process removes the nonstationary behavior in the form of regime-break.

This work studied aggregated demand models of air transportation. For future work an option comes from a model airport specific demand and then aggregate an alternative for this is the Arellano-Bond dynamic panel data, the difficulty comes from recompiling the database necessary, otherwise it should be an important resource for the planning area.

REFERENCES

Ang, A. and G. Bekaert, (2002), Regime Switches in Interest Rates, Journal of Business and Economic Statistics, 20, 163-182.

Box, G.E.P. and G.M. Jenkins (1970), Time Series Analysis; Forecasting and Control, San Francisco: Holden-Day.

Carson, R. T., et al. (2010) Forecasting (aggregate) demand for US commercial air travel. International Journal of Forecasting, doi:10.1016/j.ijforecast.2010.02.010

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction. Sixth Edition*. Chapman and Hall.

Christiano LJ (1992). Searching for breaks in GNP. J. Bus. Econ. Stat., 10, 237-250.

Chu F. (2009), Forecasting tourism demand with ARMA-based methods, Elsevier Tourism Management 30, 740–751

Cobb, C. W. and Douglas, P. H. (1928) A Theory of Production. American Economic Review, 139-65.

Cosslett, S. R. and L.-F. Lee, (1985), Serial Correlation in Discrete Variable Models, Journal of Econometrics 27, 79-97.

Dai, Qiang, Kenneth J. Singleton, and Wei Yang (2003), Regime Shifts in a Dynamic Term Structure Model of U.S. Treasury Bonds, Stanford University.

Dickey DA, Fuller WA (1979) Distributions of the Estimators for Autoregressive Time Series with a Unit Root, J. Am. Stat. Assoc., 74-366, 427-481.

Enders, W. (2004), Applied Econometric Time Series (2a ed.). John Wiley & Sons Inc.

Engle, R. F. and Granger C. W. J. (1987), Cointegration and Error Ccorrection: Representation, Estimation and Testing, Econometrica, 55, 251-276.

Febeliano, A., Müller, C., Oliveira, A.V.M. (2012), The Evolution of Air Regulation on Brazil, Aerlines Magazine e-zine edition Issue 33.

Fernandez, E., Rodriguez, R. (2010), The Causal Relationship between GDP and domestic air passenger traffic in Brazil, Transportation Planning and Technology, 33-7, 569-581.

Garcia, Rene, Richard Luger, and Eric Renault (2003), Empirical Assessment of an Intertemporal Option Pricing Model with Latent Variables, Journal of Econometrics 116, 49-83.

Goldfeld, S.M. and R.E. Quandt, 1973, A Markov Model for Switching Regression, Journal of Econometrics, 1, 3-16.

Gottman, J. M., (1981) Time-series analysis: A comprehensive introduction for social scientists. New York, NY: Cambridge University Press.

Granger, C. W. J. and Joyeux, R., (1980) An introduction to long-memory time series models and fractional differencing. Journal of Time Series Analysis, 1980, 1, 15–39.

Hamilton, J. (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica, 57-2, 357-384

Hosking, J. R. M. (1981). Fractional differencing. Biometrika, 68, 165–176.

Karlaftis M. G. (2008), Demand Forecasting in Regional Airports: Dynamic Tobit Models with Garch errors, Sitraer 7, 100-111

Kim, C. J. (1994) Dynamic Linear Models with Markov-Switching, Journal of Econometrics, 60, 1-22.

McLeod, A. I., & Hipel, K.W. (1978). Preservation of the rescaled adjusted range, 1: A reassessment of the Hurst phenomenon. Water Resources Research, 14, 491–508.

Mandelbrot, B. B. (1977). Fractals: Form, chance and dimension. San Francisco:Freeman.

Marazzo M., et al. (2009), Air transport demand and economic growth in Brazil: A time series analysis, Transportation Research Part E, 46, 261–269

Michael, P., A.R. Nobay and D.A. Peel (1997), Transaction costs and nonlinear adjustment in real exchange rates: an empirical investigation, Journal of Political Economy 105, 862-879.

Perron P (1989) The great crash, the oil price shock, and the unit root hypothesis, Econometrica, 57, 1361-1401.

Perron P (1997). "Further Evidence on Breaking Trend Functions in Macroeconomic Variables, J. Econ., 80-2, 355-385.

Profillidis, V. A. (2000) Econometric and fuzzy models for the forecast of demand in the airport of Rhodes. Journal of Air Transport Management, 6, 95–100.

Sarantis, N. (1999), Modeling non-linearities in real e_ective exchange rates, Journal of International Money and Finance 18, 27{45.

Taylor, M.P., D.A. Peel and L. Sarno (2000), Nonlinear mean-reversion in exchange rate rates: towards a solution to the purchasing power parity puzzles, International Economic Review.

Teräsvirta, T. (1994) Specification, estimation, and evaluation of smooth transition autoregressive models, Journal of the American Statistical Association, 89, 208-218.

Tong, H., (1983), Threshold models in non-linear time series analysis, Lecture Notes in Statistics, 21, Heidelberg: Springer.

van Dijk, D. and P.H. Franses (1999) Modeling multiple regimes in the business cycle, Macroeco-nomic Dynamics, 3, 311-340.

Zivot E, Andrews K (1992). "Further Evidence On The Great Crash, The Oil Price Shock, and The Unit Root Hypothesis", J. Bus. Econ. Stat., 10-10, 251–70.

Website References

Central Bank of Brazil: http://www.bcb.gov.br/?SERIETEMP

Brazilian Institute of Geopolitics and Statistics: http://seriesestatisticas.ibge.gov.br/

National Agency of Civil Aviation: http://www.anac.gov.br/Area.aspx?ttCD_CHAVE=26