# Intro

Time series forecasting, a subset of more general statistical forecasting, pertains to using mathematical models and methods to best explain the movement of a particular metric with respect to time. A model of choice is first applied to known observations, in order to *fit* the model to the data at hand; this step optimizes the model parameters to best replicate what we have already observed. Then we place our trust in model itself to forecast our metric of interest into the future. Model choice thus becomes a critical component of accurate forecasting, and is the main discussion point in this report.

NJSC consulting was tasked with forecasting monthly bankruptcy rates for Canada, for the period starting on January 2011 and ending in December 2012. Also specified in the instructions was that the forecast should be precise and accurate, and that no preference was given to any particular kind of model.

At our disposal were two files:

* “train.csv” contained monthly observations of four variables, from January 1987 to December 2010. These four variables were comprised of the Canadian national bankruptcy rate, the unemployment rate, the house price index, and the country’s population count.
* “test.csv” contained monthly observations of three variables, from January 2011 to December 2012. These three variables were comprised of the Canadian national unemployment rate, the house price index, and the country’s population count.

The data input received from the client can thus be visualized by the following diagram, where the blue bars represent observed data, while the orange bar is what we are trying to forecast.

Bankruptcy Rate

Unemployment Rate

House Price Index

Population count

Jan 2011 – Dec 2012

Jan 1987 – Dec 2010

# VAR

Another multivariate option to forecast the bankruptcy rate is to consider the variables on hand, or a subset of them, as endogenously related. This means that unlike in SARIMAX where we assume no causal effect of the response on any external variables, here we assume the response variable has a causal influence on at least one other variable to be included in the model. In such a situation, we turn to a vector auto-regression (VAR) model to assess the relationship between the response variable and all others which are endogenously related to it.

VAR models can be made more complex in the hopes of better modeling the response variable, and making better predictions. Exogenous variables can be added into the model for instance, as can seasonality indicators.

In our case, the best performing VAR model was one with bankruptcy rate and housing price index as the two endogenous variables. Despite further adding the unemployment rate as an exogenous variable to the model, as well as incorporating seasonality indicators, the values of our predictions on the validation set were not as accurate as those of other models.