Chapter 11

DETECTING SHADOW ECONOMY SIZES WITH
SYMBOLIC REGRESSION

Philip D. Truscott¹ and Michael F. Korns²

¹Department of Information Systems and Computer Science, Ateneo de Manila University, Loyola Hts, Quezon City, Philippines. ²Korns Associates, Shang Grand Tower, Unit 17E, 98 Perea Street, Makati 1229, Philippines.

Abstract

This chapter examines the use of symbolic regression to tackle a real world problem taken from economics: the estimation of the size a country’s ‘shadow’ economy. For the purposes of this chapter this is defined as a country’s total monetary economic activity after subtracting the official Gross Domestic Product. A wide variety of methodologies are now used to estimate this. Some have been criticized for an excessive reliance on subjective predictive variables. Others use predictive data that are not available for many developing countries. This chapter explores the feasibility of developing a general-purpose regression formula using objective development indicators. The dependent variables were 260 shadow economy measurements for various countries from the period 1990-2006. Using 16 independent variables, seven basis functions, and a depth of one grammar level a search space of \(10^{13}\) was created. This chapter focuses on the power conferred by an abstract expression grammar allowing the specification of a universal goal formula with grammar depth control, and the customization of the scoring process that defines the champion formula that ‘survives’ the evolutionary process. Initial searching based purely on R-Squared failed to produce plausible shadow economy estimates. Later searches employed a customized scoring methodology. This produced a good fit based on four variables: GDP, energy consumption squared, this size of the urban population, and the square of this figure. The same formula produced plausible estimates for an out of sample set of 510 countries for the years 2003-2005 and 2007. Though shadow economy prediction will be controversial for some time to come, this methodology may be the most powerful estimation formula currently available for purposes that require verifiable data and a single global formula.

Keywords: abstract expression grammars, customized scoring, grammar template genetic programming, genetic algorithms, universal form goal search
1. Introduction

This chapter describes the use of symbolic regression to tackle a research problem from economics: the estimation of shadow economy sizes.

Social scientists have long been aware of various model construction tools to select the best-fitting combination of variables using software packages like SPSS (Nurusis, 1996) and SAS (Muller, 2002). Compared to these tools, Symbolic Regression (Korns, 2010) involves a process that adds many orders of magnitude to the size of the search space typically attempted by social scientists. Korns has developed an Abstract Expression Grammar. His software for performing the search will be referred to by the acronym ARC (for Abstract Regression Classification).

Two concepts in particular add to search space size:

- **Abstract functions** Abstract Expression Grammar Symbolic Regression allows a variable to be modified by mathematical functions not defined explicitly. For example, an 'abstract' function might include the square root, square, cube, log, exponent or be left unmodified. All of these functions might be applied to the independent variables during the model fitting process.

- **Grammar Level Depth** Abstract Expression Grammar Symbolic Regression defines a generic form for the specification of models that use variables in combinations of varying complexity. To illustrate this concept consider a model that predicts economic activity from energy consumption and population. At a grammar depth of 0 each of these will be considered as separate variables within the goal formula. At a grammar depth of 1 the goal formula might include the term "energy consumption * population". At a grammar depth of 2, two functions might be applied such as "cube(energy consumption * population)".

With five basis functions and three grammar levels, the search space of potential models rises to $10^{852}$.

2. Shadow Economy Estimation Methods

There is currently no general agreement among economists on the correct procedure for measuring the size of a country’s 'shadow' or 'informal' economy. For the purposes of this paper it will be defined as a country’s total monetary economic activity after subtracting the official GDP statistic (hence referred to as the IDP - Informal Domestic Product and it includes all black market and other illegal activities including the gray activities in between).

Directly measuring the shadow economy is difficult; critical data is missing or because available data cannot be verified. Some researchers have attempted