# DETECTING SHADOW ECONOMY SIZES WITH SYMBOLIC REGRESSION

Philip D. Truscott[1] and Michael F. Korns[2]

**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 base 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

## 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 (e.g. Nurušis 1999) and SAS (Muller 2002). Compared to these tools, Korns (2010) describes 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.

[1] Department of Information Systems and Computer Science, Ateneo de Manila University, Loyola Hts, Quezon City, Philippines.

[2] Korns Associates, 1 Plum Hollow, Henderson, Nevada 89052 USA.

His software for performing the search will be referred to by the acronym ARC (what does this stand for?).

Three concepts in particular add to this size:

a) **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.
b) **Grammar Level Specification** 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 $10^{852}$.

## 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 to measure the IDP through the ‘currency demand’ method. This examines the ratio of cash holdings to bank deposits, on the assumption that informal businesses will want to evade the reporting procedures of financial institutions. Macroeconomists have built models that simulate the IDP size in relation to tax rate changes and changes in currency demand. Unfortunately this process gives no procedure for measuring the IDP size in the base year, which has lead some model builders to assume a *zero* IDP as their starting point which has been criticized as an unrealistically heroic assumption.

Kaufman and Kaliberda (1996) estimated the size of various post-Soviet economies by assuming that they grew in proportion to electricity consumption in the years after 1989. As with the currency demand approach they could not use the same methodology for their base year and had to use initial estimates from the European Bank of Reconstruction and Development.

Laćko (2000) describes a more complex electricity-based approach that takes into account such things as the country's average temperature and the cost per unit of electricity.

Hyun and Yoo (1998) use surveys of individuals in an attempt to measure the true level of economic activity by multiplying the micro-level data up to the national level. This is one of the most objective approaches in that it uses hard data and few assumptions. However its applicability is limited by the small number of countries where such survey data is available and by the difficulty of verifying the survey data.

The most extensive set of country statistics has been published by Schneider et. al. (2010) which shows IDP estimates for 162 countries. This approach uses independent variables that include subjective freedom indices from the Heritage Foundation (for example a 'Business Freedom Index' defines zero as the least business freedom and 100 as the most). Their dependent variables include a measurement of currency demand, growth rate of GDP per capita and labor force participation rate. Their methodology is described by the acronym MIMIC (Multiple Indicators Multiple Causes). The MIMIC model approach is considerably more complex than a multiple regression formula (see Schneider 2010). For micro-economists who favor hard data, this approach suffers from the defect that the quantities that are 'predicted' have not, in themselves, been proven to be measurements of the shadow economy. However the MIMIC enthusiasts can derive some comfort from the fact their IDP estimates for developed countries are often similar to those produced by other methods.

These descriptions are intended to give a brief overview to show that the current estimation methodologies are both diverse and controversial. A much fuller review of shadow economy estimations can be found in Schneider et al (2010) (who tend to downplay the value of direct measurement techniques using micro-simulation). Laćko (2000) also gives a lengthy research review which is less sanguine about the MIMIC technique (calling it the 'soft model' approach).

This paper examines the following question. Can a single global regression equation 'predict' the results of these diverse shadow economy estimates using objective data on development indicators? Currently the most extensive prediction set uses three separate estimation formulae for three groups of countries: OECD (developed) countries[*], former Soviet and allied states, and other (mainly developing) countries. The

[*] The OECD group only broadly conforms to OECD membership in 2010. Slovenia and Mexico, for example, are not placed in the developed country group though they have now acceded to the OECD. The Republic of Korea is included in the developed country group though it only joined OECD in the mid 1990s.

development of a single predictive formula would enable policy research to compare developed and developing countries more easily.

## The Dependent Variables

The dependent variables used in this analysis are estimates of the size of the shadow economy for various countries from the years 1989-2006 using different methodologies as shown in table 1 below. One of the goals of this research was to facilitate analysis on taxation policies. Therefore the country estimates taken from the MIMIC research were limited to those that do not use tax burden variables as predictors.

**Table 1: Sources of Shadow Economy Estimate**

| Estimation Method | Author | Number of Country Estimates |
|---|---|---|
| Electricity Consumption | Lackó (1996, 1997a, 1997b, 1999) | 19 |
| Currency Demand | Schneider (1994, 1998) | 30 |
| Electricity Consumption | Johnson, Kaufmann, and Zoido-Lobatón (1998a, 1998b) | 18 |
| Electricity Consumption | Schneider using data from Lackó (1996) | 25 |
| Electricity Consumption | Johnson, S., Kaufmann, D., and Zoido-Lobatón, P. (1998) | 52 |
| Electricity Consumption | Lackó, M (1999) | 53 |
| Survey Micro-data | Hyun and Yoo (1998) | 5 |
| MIMIC Model | Schneider, Buehn & Montenegro (12) | 61 |
| Various Single Country Studies | Hartzenburg, G.M., and Leimann, A. (1992)<br>Bagachwa, M.S.D. and A. Naho (1995)<br>Pozo, Susan (ed.) (1996)<br>Bhattacharyya, Dilip K. (1999)<br>Madzarevic, S and Mikulic, D., (1997) | 5 |
| Total | | 267 |

The various measurements in Table 1 may include the same country's IDP for the same year using different methodologies. The approach of collecting different estimations is similar to the technique of taking a "polls of polls" to summarize the forecasts of different economists[*].

The IDPs estimated by the research in Table 1 express the shadow economy as a proportion of the official GDP. These proportions are then used to calculate the Comprehensive Domestic Product (CDP) for each country in each year. This is defined as the official GDP plus the IDP.

[*] Where a multi-year study is quoted the last year is taken to be the year that the estimate applies to. This is because many of these multi-year studies take third party estimates for the IDP size in the base year and then refine it over time.

It is the CDP and not the IDP that is the dependent variable of the regression equations shown below.

**Table 2: Average, Minimum and Maximum Shadow Economy Sizes**

| IDP (Shadow Economy) as % of Official GDP | | | | | |
|---|---|---|---|---|---|
| Region | Mean | Minimum | | Maximum | |
| 1 (OECD) | 14.8 | 5.1 | (Austria 1990) | 50.5 | (Spain 1990) |
| 2 (Mainly Ex-Soviet) | 29.0 | 6.4 | (Czech Rep. 1990) | 74.9 | (Russia 1995) |
| 3 (Mainly Developing) | 34.9 | 9.0 | (South Africa 1990) | 76.0 | (Nigeria 1990) |

Table 2 gives a general impression of the range of shadow economy sizes that have been estimated by the different techniques.

## The Independent Variables

The independent variables are various development indicators that to try to predict the size of the CDP objectively. Not only electricity but total energy usage increases as economies grow. More developed economies also tend to be more urban and the birth rate tends to decline. More developed economies are less agricultural and more densely populated. This methodology does not seek to predict economic evasiveness directly. This contrasts with approaches like the currency demand method which measures the amount of cash kept out of banks. The current technique tries to measure total economic activity. The IDP is estimated afterwards by subtracting the official GDP.

For certain countries and years the development indicator data in Table 2 were missing. After deleting these data, 260 country estimates remained with all the corresponding development indicators for the years concerned.

**Among the development indicators in**

Table 3 the power consumption variables (per capita electricity, total electricity, and total energy) are particularly interesting to researchers because the 1990s showed them to be powerful predictors of both economic growth and contraction; the former soviet states showed absolute *declines* in electricity consumption *and* GDP. Western Europe and North America showed increases in both. The telecommunications variables covering mobile phones and broadband Internet subscriptions have been combined with landline phones to create two composite variables:

- COMSTWO includes two types of telephone fixed line and Mobiles
- COMSTHREE includes both types of Phone and Broadband Internet subscriptions

**Table 3: Development Indicators used as IDP predictors**

| | |
|---|---|
| URBAN_PER | Proportion of population that is urban |
| URBAN_POP | Number of urban persons |
| LABOR_PARTICIPATION | Labor participation rate, total (% of total population ages 15+) |
| BIRTH_RATE | Birth rate, crude (per 1,000 people) |
| ELEC_PER_CAP | Electric power consumption (kWh per capita) |
| TOTELEC | Total Electric power consumption (kWh) |
| ENERGY | Indicator: Energy use (kilotons of oil equivalent) |
| C02 | CO2 emissions (kilotons) |
| PHONES | Telephone lines |
| IMPORTS | Imports of goods and services (current US$) |
| POPULATION | Population |
| POPDENS | Population density (people per sq. km) |
| ARABLE_PER_AG | Arable Land as a % of Agricultural Land |
| AGLAND_PER | Agricultural land (% of land area) |
| ARABLE_PER | Arable land (% of land area) |
| AREA | Surface area (sq. km) |
| COMSTWO | Telephone (landlines) and Mobile Phone subscriptions |
| COMSTHREE | Telephones (landlines), Mobiles + Broadband Subscriptions |
| GDP_PPP | GDP, PPP (current international $) |

Combining the newer communication methods with landline phones avoids a dataset with many missing values for the 1990s.

The ability to predict shadow economy sizes with objective development indicators has important policy research implications. Global banking institutions might wish to assess if particular types of tax increase the size of the IDP. Are direct taxes more damaging than sales taxes? Are taxes on starting a business or employing workers more negative in their effect on tax revenues? Since the MIMIC approach uses tax variables to predict the IDP, it is impossible to make such analyses with their data. Such analyses would involve making tax data both the independent and dependent variables. For developing countries tax collection is often extremely inefficient in turn leading to poor quality education, health care and environmental protection. The inability of developing country governments to maintain law and order, and impose their will generally led them to be termed ‘soft states’ by the Nobel Prize winning economist Myrdal (1968). A fuller understanding of the causes and cures of large shadow economies is essential for addressing this problem.

## Model Optimization with Standard Scoring

Initial testing of Abstract Expression Grammar Symbolic Regression showed that the usual method of assigning a score to potential champion formulae was unsuitable for the task at hand. Previous papers using this approach have based the scoring on a combination of Normalized Least Square Error (NLSE) and Tail Classification Error (TCE) which measures how well the regression champion classifies the bottom 10% and the top 10% of the data set (Korns 2010).

Using five variables and one grammar level of depth the champion formula estimated the Comprehensive Domestic Products with an R-squared of over 96%. However when the official GDP figures where subtracted from the CDP figures, some countries had negative values. Common sense indicates that no populations are so honest that they have zero-sized informal economies. Negative IDPs are also unreasonable.

## Customized Scoring

One of the more powerful features of ARC is the ability to fine tune the scoring that determines which formulae will survive the evolutionary process. In the case of IDP estimation a scoring methodology was required to penalize potential champions with negative IDPs as well as those with excessively high IDPs.

The approach adopted was to calculate the IDP for each of the 260 country measurements for each potential champion formula. The official GDP was stored in a global array for each country and measurement year. The official GDP was then subtracted from the predicted Comprehensive Domestic Product for that country/year for the regression formula being scored. The result was the 'predicted' Informal Domestic Product. In order to penalize implausible predictions, a given formula's fitness score was multiplied by 1.25 if:

- One country's estimated IDP% was lower than 3%
- One country's estimated IDP% was higher than 140%

In theory the scoring process would have penalized any formula that produced an IDP estimate higher than 100%. The largest IDP% estimate in Table 2 is 76% for Nigeria in 1990. Philosophically there is no reason to discount the possibility of very large shadow economies. An IDP percentage above 100% simply means that the shadow economy is larger than the official GDP estimate. For countries recently emerging from communism with sharply contracting economies (like some central Asian states in the 1990s) such large shadow economies seem entirely plausible.

Initial tests of this procedure found that the scoring penalties prevented the search algorithm from finding any strong champions at all. This problem was solved by allowing the NLSE-based scoring to proceed in the normal way until an NLSE value lower than 0.2 was discovered. Only at that point was the scoring modified according to the minimum and maximum shadow economy size.

The customized scoring function could he adapted and reprogrammed outside the main calculation engine using a version of LISP (whose details are described at www.korns.com/Document_Lisp_Language_Guide.html). The programming required would seem reasonably familiar to any researcher with familiarity with a fourth generation language, though the use of pre-fix operators takes some getting used to for those more accustomed to in-fix operators.

## Salient Language Features

For the purposes of the current research the most important features of the Abstract Expression Grammar were as follows:

- Abstract Function Specification
- Basis Function Limitation
- The Universal Goal Formula
- Grammar Level Depth Control

This paper aims to provide a practical example of the Abstract Expression Grammar in use. For a more general overview see Korns (2010).

## The Universal Goal Formula

The Abstract Expression Grammar allows the user to define the general form of the desired regression equation with a LISP statement illustrated by the following example:

```
(E1):   universal(2,4,v)
```

This example specifies that the desired formula should be "universal" in form.  The three parameters following the keyword universal constrain the formula as follows:

- **Grammar Level Depth**: 1st parameter is the permitted grammar level depth to be searched (the example shown in E1 above searches to a depth of two grammar levels). A zero in this position would produce the simplest formulae. Since the independent variables are unmodified the zero grammar level makes the process resemble model testing in SAS and SPSS. Increasing this value greatly adds to the size of the search space.
- **Number of Terms**: The 2nd parameter specifies the number of terms that will be used as independent variables. Since the example in E1 above limits the predictors to four terms and the final parameter is set to 'V' this means the goal formula will have exactly four independent variables.
- **Term Format:** The 3rd parameter specifies whether the terms should be in the form of a variable or an abstract term. In E1 above the letter 'V' has been specified indicating the use of variables. If the letter 'T' had been used an abstract term could be entered (which means that a variable could be replaced by a constant).

In the interests of methodological rigor the coefficients of each ARC champion formula were recalculated in SPSS version 15. The coefficients shown in the tables below are those produced by SPSS. The significance scores are those produced by SPSS using the Student's T procedure.

## Baseline Test: Searching at Grammar Depth Zero

A baseline test was conducted which forced ARC to search in a way that imitated the model testing procedures available in traditional statistical packages like SAS and SPSS. At a grammar depth of zero using variable terms (rather than abstract terms) ARC tests different combinations of variables unmodified by each other or by mathematical functions.

```
(E2):   universal (0,5,v)
```

Equation E2 above specifies that the desired goal formula should have five unmodified variable terms.

The resulting shadow economy estimates immediately show the need for customized scoring. Even though all of the terms were significant at the 1% level and an R-Squared of 0.95 was reached some shadow economies were implausibly high or low. As can be seen from Table 4 Trinidad[*] was estimated to have a negative shadow economy, while Mongolia had one of 230%.

**Table 4: Champion Formula - Grammar Depth 0**

| Region | Mean | Smallest Shadow Economy | Largest Shadow Economy |
|---|---|---|---|
| 1 OECD | 16.6% | 9.7% (USA 1990) | 35.0% (Australia 1990) |
| 2 Ex-Soviet | 27.4% | 9.5% (Ukraine 1990) | 96.6% (Georgia 1995) |
| 3 Developing | 37.7% | -3.1% (Trinidad 2006) | 230.3% (Mongolia 2006) |

| Variable | Beta | t | Significance |
|---|---|---|---|
| GDP_PPP | 1.113 | 91.643 | 0 |
| URBAN_POP | 3007.583 | 9.506 | 0 |
| CO2 | -122862.124 | -8.398 | 0 |
| AREA | 9329.823 | 6.395 | 0 |
| POPULATION | -577.44 | -6.28 | 0 |

## A Search using Grammar Depth 1

Searching at a grammar depth of 1 allowed ARC to use mathematical functions and operators. Where the operators take two operands this caused ARC to use some very complex composite terms. The first term below uses the addition operator to combine the official GDP with the urban population. Those who express skepticism about elaborate model searching procedures would immediately raise the problem of 'over-

[*] The official name is the Republic of Trinidad and Tobago.

fitting'. Why should the sum of the official GDP and the urban population be an indicator of the Comprehensive Domestic Product? Re-running the champion formula in SPSS gives some credence to the over-fitting objection in this case because one of the five terms was removed because it failed a built-in test for co-linearity. So SPSS constructed a regression formula without the last term which was the sum of Energy consumption and COMSTWO (Landline Phones + Mobile Phones). As with the previous champion formula, high significance levels and a high R-Square values did not guarantee plausible estimates (see Table 5).

It should be pointed out that many academic papers publish regression formulae using two variables in combination (such as GDP per capita). The champion shown in Table 5 was found after only 6.24 thousand Well Formed Formulas (WFFs). If a larger number of WFFs had been searched, ARC might have discovered a better fit using more complex (and more plausible) binary terms. For the purposes of this paper, more parsimonious model selection was attempted rather than a longer search process.

**Table 5: Champion Formula - Grammar Depth 1**

| REGION | Mean | Smallest Shadow Economy | | Largest Shadow Economy | |
|---|---|---|---|---|---|
| 1 OECD | 25.1% | 10.5% | (USA 1990) | 57.2% | (Ireland 1990) |
| 2 Ex-Soviet | 38.1% | 11.9% | (Russia 1995) | 89.7% | (Estonia 1995) |
| 3 Developing | 49.5% | 0.2% | (Vietnam 2006) | 141.4% | (Togo 2006) |

| Variable | Beta | t | Significance |
|---|---|---|---|
| GDP + URBAN_POP | 1.16 | 95.966 | 0 |
| SQRT NET_IMPORTS | 139026.804 | 6.125 | 0 |
| PHONES + COMSTWO | -5196.454 | -8.088 | 0 |
| COMSTWO + ENERGY | 7074.87 | 8.393 | 0 |
| Note : The term (GDP_PPP + COMSTWO) was removed from this champion because it failed a built-in SPSS co linearity check. | | | |

## Searching with Function Limitations

Four of the five terms in Table 5 use the plus operator to construct complex terms involving two variables. The resulting complex formula illustrates a common criticism of regression formulae emerging from genetic programming – the so-called 'bloated' nature of the winners. Luckily ARC provides several effective ways to limit bloat. One of them is to apply a restriction to the function that can be selected during model evolution.

```
(E3):  universal(2,5,v)
           where {op(noop,sqrt,square,cube,log,exp,abs)};
```

Equation E3 applies a “where” clause to the goal search formula that restricts the search to unary mathematical operators. At the same time the first parameter has been set to “2” which raises the grammar depth by one level. This allows a variable to be modified by two functions in combination. The first term shown in Table 6 used two operators that cancelled each other out (the square root of GDP squared).

Four of the five other terms used the ABS function which would have removed the minus sign from any negative variable values. Since there were no negatives, this function had no effect. In a sense the search algorithm appears to be using the ABS function as a “No Operator” function. This was equivalent to ARC choosing keyword “noop” in equation E3.

**Table 6: Champion Formula - Unary Operators & Grammar Depth 2**

| REGION | Mean | Smallest Shadow Economy | | Largest Shadow Economy | |
|---|---|---|---|---|---|
| 1 OECD | 21.8% | 9.4% | (USA 1990) | 31.3% | (Ireland 1990) |
| 2 Ex-Soviet | 42.1% | 14.7% | (Ukraine 1990) | 133.8% | (Kyrgyz Rep. 1995) |
| 3 Developing | 34.5% | 9.9% | (China 2006) | 123.6% | (Togo 2006) |

| Variable | Beta | t | Significance |
|---|---|---|---|
| SQRT_GDP$^{2}$ | 1.188 | 123.642 | 0 |
| ABS ENERGY$^{2}$ | -0.289 | -5.177 | 0 |
| ABS AREA$^{3}$ | 6.04E-11 | 9.474 | 0 |
| ENERGY$^{6}$ | 9.22E-27 | 2.819 | 0.005 |
| ABS LOG GDP | 232909816.1 | 1.472 | 0.142 |

The third term (AREA to the power of 6) is the equivalent of repeating the “cube” function for area: “cube(cube(AREA))”. In natural language one could express this as “the cube of area cubed”.

As with Table 5, the champion formula at grammar depth two might have been improved by allowing more search time. Also some of the individual country estimates were implausible (such as the Irish estimate of over 31%). The final term (the Log of GDP) was not significant at the 10% level.

## Function Restriction at Grammar Depth 1

In the interests of parsimony the function-restricted goal formula was further restricted to 1 grammar level of depth. This only required a small change to the goal formula. The first parameter was reduced to 1, resulting in this goal:

```
(E4):  universal(1,5,v)
            where {op(noop,sqrt,square,cube,log,exp,abs)};
```

This search was allowed to run for 30.73K WFFs but ARC found the eventual champion after only 5.78K WFFs. Using both customized scoring and a relatively parsimonious goal formula ARC produced the first of a series of plausible estimation sets.

**Table 7: Champion Formula - Unary Operators & Grammar Depth 1**

| Region | Mean | Smallest Shadow Economy | Largest Shadow Economy |
|---|---|---|---|
| 1 OECD | 14.9% | 9.5% (USA 1993) | 19.1% (Korea 1996) |
| 2 Ex-Soviet | 38.3% | 14.7% (Slovenia 1995) | 95.1% (Georgia 1995) |
| 3 Developing | 35.7% | 7.6% (Trinidad 2006) | 113.8% (Togo 2006) |

| Variable | Beta | t | Significance |
|---|---|---|---|
| PHONES$^3$ | -1.4E-13 | -7.701 | 0 |
| COMSTHREE$^2$ | 0.00000943 | 7.537 | 0 |
| POPULATION$^3$ | -4.95E-16 | -10.012 | 0 |
| URBAN_POP | 2135.92 | 13.397 | 0 |
| GDP | 1.063 | 135.62 | 0 |

The countries that commonly have the smallest shadow economies in the existing research (Switzerland, Austria and the USA) have similarly low estimates using the formula in Table 7. Togo is the only country estimate above 100%. The next highest was Georgia with an estimate of 95%. Trinidad's score might seem to be unduly low, but a similar low estimate is given by Schneider (2010). If one defines 'free' business registration as a process requiring less 1.5% of per capita incomes then Trinidad is virtually the only non-OECD country in the free registration category[*], which lends plausibility to the idea that it might have a small shadow economy (since its government makes it so easy for entrepreneurs to be honest reporters).

A regression champion was found similar to the formula at the bottom of Table 7 that used total energy consumption in place of the communications variables. This is shown in Table 8 below.

[*] World Bank, (2010), "Doing Business 2010", accessed on March 31st 2010 from <http://www.doingbusiness.org/ExploreTopics/StartingBusiness/>

**Table 8: 'Energy' Champion - Unary Operators & Grammar Depth 1**

| Region | Mean | Smallest Shadow Economy | | Largest Shadow Economy | |
|---|---|---|---|---|---|
| 1 (OECD) | 15.6% | 9.8% | (USA 1993) | 19.3% | (Korea 1996) |
| 2 (Ex-Soviet) | 40.3% | 15.9% | (Slovenia 1995) | 100.5% | (Georgia 1993) |
| 3 (Developing) | 30.0% | 8.3% | (Trinidad 2006) | 120.2% | (Togo 2006) |
| | | | | | |
| Variable | | Beta | | t | Significance |
| ENERGY$^3$ | | -1.94E-08 | | -2.469 | 0.014 |
| GDP_PPP | | 1.071 | | 88.924 | 0 |
| URBAN_POP | | 2255.715 | | 11.57 | 0 |
| URBAN_POP$^2$ | | -2.72E-06 | | -7.839 | 0 |

## An Out of Sample Check

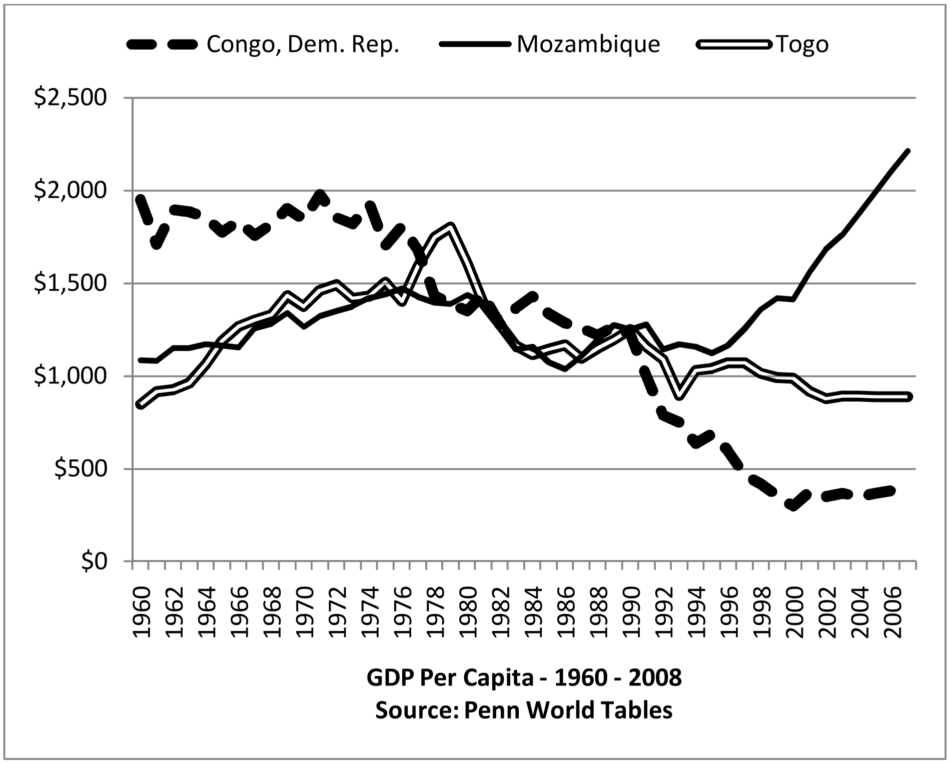
One of the most commonly suggested solutions to over-fitting is to calculate regression estimates for a set of cases separate from the original data. The results of an out of sample check are shown in Table 9 below. The GDP, energy consumption and urbanization values for the years 2003-2005 and 2007 were applied to the four term regression formula shown in Table 8.

**Table 9: Out of Sample Prediction for 510 country estimates in 2003-05 & 07**

| Region | Average | Smallest Shadow Economy | Largest Shadow Economy |
|---|---|---|---|
| 1 (OECD) | 12.3% | 8.1% (USA 2007) | 15.6% (Korea 2003) |
| 2 (Ex-Soviet) | 27.6% | 11.2% (Slovenia 2007) | 62.8% (Moldova 2003) |
| 3 (Developing) | 41.2% | 8.3% (Trinidad 2007) | 305.8% (Congo, D.R. 2003) |
| 510 shadow economy predictions for various countries in 2003-05 & 2007 based on official GDP, Urban Population, Urban Population Squared, and Energy Consumption cubed as shown in table 8 above (2006 data was “in-sample” and so has been omitted). | | | |

In Table 9 the Democratic Republic of Congo is a notable outlier with an IDP of 305%. However this was one of only four countries with estimates higher than 100%. The others were Mozambique (Highest 140%), Togo (127%), Ghana (107%) and Haiti (101%). Given the political pressures on the GDP estimation procedure in developing countries it seems entirely plausible that a broadly accurate estimation formula will produce such outliers in a given year. In the developing world a government economist may be pressured to show unduly high economic growth to secure government loans from global financial institutions. If tax revenues are low in relation to official GDP the opposite pressure might apply. Tanzania was once penalized by a sharp reduction in foreign aid due to ‘under-taxation.’ (Gould 2005).

**Figure 1: GDP Per Capita in Togo, Mozambique and Dem.Rep.Congo**



**GDP Per Capita - 1960 - 2008**
**Source: Penn World Tables**

An accurate estimation formula based heavily on urbanization may correctly predict the CDP while economies are growing but not when there are sharp economic contractions. Populations that move to the cities will not flood back to the countryside in a recession. The three largest IDPs mentioned above were all in countries that experienced significant economic contractions between 1960 and 2007 (see figure 1). Togo's per capita income stayed about the same over this period while the per capita incomes in the rest of the world increased from $4,200 USD to $13,600 (Heston, Summers and Aten 2006). The Democratic Republic of the Congo actually experienced a sharp decline. A predictive formula relying more heavily on energy data might capture such declines, but it seems unlikely that a global formula solely based on energy could have a high level of explanatory power with the data currently available. Given the intense war and civil disturbances in the countries shown in figure 1, the high IDP estimates shown may in fact be true values even though they exceed those in previously published research. As Schneider (2010) points out this type of research involves a scientific passion for "knowing the unknown."

Researchers who want to make detailed year-to-year predictions of IDP will not be satisfied with these estimations. It is not claimed that the predictor variables shown can track both downward and upward movements in the size of the CDP in a given year, however over a 5-10 year period the estimates shown may be the more accurate than the MIMIC estimates. For researchers who prefer predictions based on 'harder' quantitative data than the MIMIC approach, the regression formulae in Tables 7-8 may currently provide the most accurate procedure for the countries covered by the formula shown in Table 8 (126 countries with known energy, GDP and urbanization data).

## Assessing ARC's Abstract Expression Grammar

ARC revealed some missing analytical tools during the conduct of this research. It would be useful to use Student's T statistics and co-linearity values calculated inside ARC. Given ARC's main strengths, we recommend adding these additional statistical tools as soon as possible. ARC was able to find useful regression champions within a gigantic search space, while supporting the very high degree of user control necessary in real world applications. Even with function restriction, four variables, and a grammar depth of 1 a search space of 100 trillion combinations was created. This is such a large area that traditional statistical packages would have been entirely impractical for these estimations. The Abstract Expression Grammar shown in equations E1-E4 allowed the parsimony of the goal formula to be adjusted quickly with a trivial amount of effort. The regression formulae shown above are a useful gain in knowledge that would have been difficult to achieve without a genetic programming approach to regression analysis.

## References

Bagachwa, M.S.D. and A. Naho (1995): "Estimating the Second Economy in Tanzania," World Development, 23:8, pp. 1387-1399.

Bhattacharyya, Dilip K. (1999): "On the Economic Rationale of Estimating the Hidden Economy," The Economic Journal, 109:456, pp. 348-359.

Gould, J., (2005), "The New Conditionality: the politics of poverty reduction strategies", London: Zed Books.

Hartzenburg, G.M., and Leimann, A. (1992): "The Informal Economy and its Growth Potential," in E. Adebian and B. Standish. (eds.): Economic Growth in South Africa. Oxford: Oxford University Press, pp. 187-214.

Heston, A., Summers, R., and Aten, B., (2006) Penn World Table Version 6.2, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania. Accessed on April 6th 2010 from <http://pwt.econ.upenn.edu/php_site/pwt_index.php>

Johnson, S., Kaufmann, D., and Zoido-Lobatón, P. (1998a): "Regulatory Discretion and the Unofficial Economy," The American Economic Review, 88:2, pp. 387-392.

Johnson, Simon; Kaufmann, Daniel and Pablo Zoido-Lobatón (1998b): Corruption, Public Finances and the Unofficial Economy. Washington, D.C.: The World Bank, discussion paper.

Kaufmann, D., Kaliberda, A. (1996), Integrating the Unofficial Economy into the Dynamics of Post Socialist Economies: A framework of Analyses and Evidence. In B. Kaminski (ed.), Economic Transition in Russia and the New States of Eurasia, London: M.E. Sharpe, pp.81–120.

Korns, Michael F., (2010), Abstract Expression Grammar Symbolic Regression, in Riolo, Rick L., O'Reilly, Una-May, and McConaghy, Trent, editors, Genetic Programming Theory and Practice VIII, Springer, Ann Arbor.

Lackó, M (1997a): The Hidden Economies of Visegrád Countries in International Comparison: A Household Electricity Approach. Hungary: Institute of Economics, working paper.

Lackó, M (1997b): Do Power Consumption Data Tell the Story? (Electricity Intensity and the Hidden Economy in Post-Socialist Countries). Laxenburg: International Institute for Applied Systems Analysis (IIASA), working paper.

Lackó, M (1998): "The Hidden Economies of Visegrad Countries in International Comparison: A Household Electricity Approach." in L. Halpern and Ch. Wyplosz (eds.): Hungary: Towards a Market Economy. Cambridge (Mass.): Cambridge University Press, p.128-152.

Lackó, M (1999): Hidden Economy an Unknown Quantity? Comparative Analyses of Hidden Economies in Transition Countries in 1989-95. Working paper 9905. Department of Economics, University of Linz, Austria.

Laćko, Maria, (2000), Hidden economy - an unknown quantity? Comparative analysis of hidden economies in transition countries, 1989-95, in Economies in Transition, Wiley-Blackwell, Vol 8 (2000), 117-149.

Madzarevic, Sanja and Davor Mikulic (1997): Measuring the unofficial economy by the system of national accounts, Zagreb: Institute of Public Finance, working paper.

Muller, K.E., (2002), Regression and ANOVA: an integrated approach using SAS software, Cary NC: SAS Publications.

Myrdal, G., (1968), Asian drama; an inquiry into the poverty of nations, New York: Twentieth Century Fund.

Nurušis, M.J., (1999), SPSS regression models 10.0, Chicago IL: SPSS Inc.

Pozo, Susan (ed.) (1996): Exploring the Underground Economy: Studies of Illegal and Unreported Activity. Michigan: W.E. Upjohn, Institute for Employment Research.

Schneider, F., (1994): "Measuring the Size and Development of the Shadow Economy: Can the Causes be Found and the Obstacles be Overcome?" in Hermann Brandstaetter and Werner Güth (eds.): Essays on Economic Psychology, Berlin: Springer, pp. 193-212.

Schneider, Friedrich (1998a): "Further Empirical Results of the Size of the Shadow Economy of 17 OECD-Countries over Time," Paper presented at the 54th Congress of the IIPF Cordoba, Argentina and Discussion Paper, Department of Economics, University of Linz, Linz, Austria.

Schneider, F., Buehn, A., Montenegro, C., (2010): Shadow Economies All over the World - New Estimates for 162 Countries from 1999 to 2007, The World Bank Development Research Group: Washington DC.

Yoo, I., Hyun, J.K., "International Comparison of the Black Economy: Empirical Evidence Using Micro-level Data", Korea Institute of Public Finance, Working Paper 98-04, 1998.