MDG Data Analysis Report

Author: Kobi Abayomi

INTRODUCTION/SUMMARY 3

DATA QUALITY ASSESSMENT 4

MISSINGNESS 4
STATISTICS FOR MISSINGNESS 5
TREND 8
STATISTICS FOR TREND 8
STATIONARITY 9
STATISTICS FOR STATIONARITY 9
MONOTONICITY 10
STATISTICS FOR MONOTONICITY 10
EXAMPLE I 10
EXAMPLE II 13

UNIVARIATE-BASED PREDICTION OF MDG PROGRESS 17

EXAMPLE I 17
EXAMPLE II 22

MULTIVARIATE-BASED PREDICTION OF MDG PROGRESS 25

EXAMPLE I 25
EXAMPLE II 27

SUMMARY 29

FIGURES 30

FIGURE 1 30
FIGURE 2 31
FIGURE 3 32
FIGURE 4 33
FIGURE 5 34
FIGURE 6 35
FIGURE 7 36
FIGURE 8 37
FIGURE 9 38
FIGURE 10 41

TECHNICAL APPENDIX 42

DATA QUALITY ASSESSMENT 42
MISSINGNESS 43
TREND 44
STATIONARITY 45
MONOTONICITY 46
UNIVARIATE PROGRESS PREDICTIONS 46
MULTIVARIATE PROGRESS PREDICTIONS 48

REFERENCES 50

MDG Data Analysis   October 2014

Introduction/Summary

This report is a guideline for statisticians and statistical offices in the countries that produce the data for the Millennium Development Goals. In this way, this report illustrates methodology for data quality assessment and prediction of the variables that measure the Millennium Development Goals

I am following the organization of the Millennium Development Goals 2014 Progress Chart[1] by grouping the globe into five regions:

- Africa
- Asia
- Oceania
- Latin America and the Caribbean
- Caucasus and Central Asia

and within each region I chose representative countries to illustrate the methodology

- Africa: Mauritania and Angola
- Asia: South Korea, **Thailand**, Malaysia, Pakistan
- Oceania: New Zealand
- Latin America and the Caribbean: **Ecuador** and Trinidad
- Caucuses and Central Asia: Georgia, Iran and Afghanistan.

Table 1

The countries in **bold** have in depth methodological illustrations as examples.

This report begins with the **data quality assessment** – an illustration of data availability and timeliness. This is a guide to the sort of data inspection that country level statisticians and investigators must use before attempting any data modeling. I consider the data quality assessment *a priori* modeling: the plots and charts here are – in a very direct way – are the data in a way that precedes any subsequent prediction.

The middle section of the report are the **univariate goal predictions**: these work by viewing the data as yearly repeated measurements in a time series approach.

[1] http://mdgs.un.org/unsd/mdg/Resources/Static/Products/Progress2014/Progress_E.pdf

The next section of the report is the illustration of **multivariate goal predictions**. Here the each of the MDG variables are viewed as jointly distributed with other data: both MDG and non-MDG.

In each of these approaches to goal prediction the **data are random quantities** and the modeling is tantamount to accounting for error in univariate measurement and conditional expectation. Thus, the output of these methods are point *and* interval estimates for the MDG variables.

Relevant tables are included in the body of the document – the remainder of the figures are in the suffix.

Practitioners can and should follow the methods here, in the order here. The plots in this report are methodology as well. For instance, the missingness and time-scatter plots in the data quality section below should always be precursors to modeling/prediction. The illustrations point out what is reasonable, not just what is statistically feasible.

Data Quality Assessment

The first task is to organize the data as observations – the units of analysis or cases – and variables – the categories on which measurements are taken. The data are taken from UNSTATS website[2]: the cases here are the twelve countries listed in Table 1; the variables are 81 Millennium Development Goal variables.[3]

Each of the variables has a maximum of 25 measurements at each country – these should be seen as the progression of a *time series*. In this way the data are instantiations of the progress (or lack of) towards the MDGs and **not** as repeated measurements of a steady state. This perspective guides the modeling/prediction below; as a first principle the raw data should **not** be seen as draws from a stationary process.

Missingness

Figure 1 (below in Figures) illustrates the missingness – or pattern of missing/unobserved values – in the representative data set. Darker colors are less missing.

[2] http://mdgs.un.org/unsd/mdg/Default.aspx

[3] There are a maximum of 171 variables for each of the countries in the UNSTATS data, 81 of which are coded "Yes" as being MDG.

Two countries – Trinidad and Iran – are completely missing in all 81 of the MDG variables, i.e. across all 25 years for each variable. Some variables have less missing values – for example: Internet use, number of cell phones, and tuberculosis incidence are less missing for many countries and years. In general the rate of missingness in the data is high.

Figures 2 and 3 are *index plots* (the x-axes are the available years from 1990-2014) plots of each of the MDG series: each plot-page (of 81 in total) is the data for each of the 12 representative countries at each variable.

Take Figure 2, the percentage literacy rates of 15-24 year olds: South Korea, Trinidad, New Zealand and Afghanistan have no observations at all. Thailand has 3 observations; Pakistan has 7; Ecuador has 8. These plots illustrate missingness, trend in each series, as well as a naked comparison across countries.

In Figure 3 - Consumption of an Ozone-Depleting Substance in Metric Tons – the data are available for most countries. Inspection of the plots also suggests that after an initial spike in the early 1990s, the values of the series are decreasing – quite similarly – for each of the representative countries.

Statistics for Missingness

The plots in Figure 1-3 are statistics (simply functions of data) for missingness and are the first step in modeling here. Figure 1 illustrates the entire data and Figures 2 and 3 are used to inspect individual series.

We do not make any predictions for series with completely missing data; series with more than 80 percent missing are difficult to model and make predictions for as well.

The MDG data are highly missing: overall there are 71.5 percent missing values across the 80 countries, 81 variables and 25 years. See Table 2.

Country	% Missing
Mauritania	67
Angola	71
SKorea	62
Thailand	63
Malaysia	64
NewZealand	62
Ecuador	65
Trinidad	100
Georgia	100
Afghanistan	74

Table 2

Table 3 and Table 4 are the 10 least and most missing variables across countries

Variable/Series	% Missing
Mobile Cell Subscriptions/100 people	20
Fixed Cell Subscriptions/100 people	22
Children under 5 mortality rate	23
Infant Mortality rate	23
Tuberculosis prevalence/100,000 people (midpoint)	23
Tuberculosis death rate per year per 100,000 population (midpoint)	23
Tuberculosis incidence rate per year per 100,000 population (mid-point)	23
Children 1 year old immunized against measles, percentage	24
Proportion of the population using improved drinking water sources, total	24
Tuberculosis detection rate under DOTS, percentage (mid-point)	28

Table 3

Variable/Series	% Missing
Condom use at last high-risk sex, 15-24 years old, women, percentage	100
Condom use at last high-risk sex, 15-24 years old, men, percentage	100
Proportion of fish stocks within safe biological limits	100
Proportion of species threatened with extinction	100
Total number of countries that have reached their HIPC decision points and number that have reached their HIPC completion points (cumulative**)**	100
Pct. Pop. with access to essential drugs	100
ODA received as pct. of GDI	100
Avg. imposed tariff on agriculture imposed on developing countries	100
Avg. imposed tariff on textiles imposed on developing countries	100
Avg. imposed tariff on clothing imposed on developing countries	100

Table 4

While the data are highly missing, there does appear to be a pattern of availability across countries. We restrict further modeling to variables that have less than 80 percent missing values and we choose not to impute data. In the best case variables with only 5 observations out of 25 would be observed in the final five years. Still, the error in a prediction on just these few values would be large.

See Figure 4 in the Appendix: the variables that meet the 20 percent cutoff in observed data seem to be similar across countries. Figure 5 is a frequency distribution of missingness in the variables: many variables are completely missing. Figures 6-10 (in the Appendix) illustrate the number missing per variable, across countries.

Trend

When we consider movement towards the MDGs in the data we can immediately associate this with a hypothesis test for *trend* – which we can measure as a change in mean over time. In the most general way, we can consider a null hypothesis of no significant change in the mean and the alternative when there is evidence of a change. In practice, these are often a test for a non-zero rate of change – or slope – for a particular choice of function. We consider the functionalization to be less important here; the particular hypothesis tests we choose to assess trend reflect that.

Statistics for Trend

We test for trend using a *modified binary segmentation* procedure: essentially we partition each time series and test the possibility that the mean is not constant over the series by comparing the *likelihood ratios*. If the mean significantly differs on a partition the endpoint of the partition is the *change point*. The estimator for the means is calculated on each partition only.

The intervals we use vary from series to series and country to country: since the data are very missing we can't partition in the same way every time. The maximum number of segments, though, can be fixed: as we've set the minimum number of points for estimation at 5, we look for at most 4 change points for every series.

An alternate test of trend would be to impose a linear model and then look for significance in the slope coefficient. We don't do that here for a few reasons:

- We do use a linear random effects model later in the univariate and multivariate predictions.
- In a sense the test of the coefficient in the linear model is not a just test of trend but of linear fit. Consider data that follow a polynomial or sinusoidal pattern – depending upon the degree of the polynomial or periodicity of the sinusoid a coefficient test will fail despite obvious trend. At the same time, for a linear *piecewise* model (the random effects model we do fit) defined on the segments between the change points, the trend *is* the slope. We will use this fact for a test of monotonicity, where we must impose a functionalization of the series.
- The estimators of the means and variances we generate via change point analysis will guide the random effects modeling (for at least the variance of the intercept) later.

Lastly, the change point approach allows us to generate estimators for the mean and variance within and across segments identified by the change points. These allow us test for trend, stationary and monotonicity *en banc*: from one battery of statistics. Given the accuracy of the change point procedure, this is a way to characterize the series with minimal additional modeling assumptions

See the Appendix for further elucidation.

Stationarity

The distinguishing feature of the way we model these data, as time series, is that we discard the ordinary independent, identically distributed assumption for some version of intertemporal dependence. In this way, the statistical modeling here will reflect the real world: the MDG data are measurements from persistent, temporally dependent processes – processes that should represent the human efforts to meet the MDGs.

A time series is *stationary* if the underlying process that generates it (represented by its *probability distribution*) does not vary over time. This assumption is too strong to be reasonable for the processes that generate the MDG data. At the very most we will assume the data are *weakly stationary* – in that the generating processes remain stable though they may shift mean, variance, etc.

We do depart from ordinary time series procedures somewhat because of:

- missingness: most of these series have only a fraction of data available across the entire time span (1990-2014)
- unequal intervals: as well the years between measurements differ from series to series

So we do not consider the usual preliminary time series methods: estimation of auto-covariance and auto-correlation functions, auto-regressive/moving average models. These types of estimators would impose assumptions upon the data that cannot be met here.

Statistics for Stationarity

We test for stationarity using the same *modified binary segmentation* and *likelihood ratio* test statistic after determining the change points for the mean. This allows us to test just the constancy of the variance though with a reduction in degrees of freedom from the plug in estimation of the mean.

Again, the estimators we generate from this procedure will guide our random effects model for prediction below.

Monotonicity

A time series is *monotonic* if its rate of change does not change sign. Since we are choosing to make the functional choice less important – we are modeling the data as draws from a parameterized distribution – we capture monotonicity as a change in the parameters of a distribution and not as a higher order functional form.

Statistics for Monotonicity

We use the output from the change point procedures to test and illustrate monotonicity. Simply, we use the estimators of the mean and the change points to compute slope estimates and use these to classify trend on a segment as positive, negative, or zero (+,-,0).

A change in trend from segment to segment is evidence of non-monotonicity in the series.

Example I

Let's take the data series for Population below $1 (PPP) per day, percentage for Thailand as an example. Out of a possible twenty-five observations thirteen are missing.

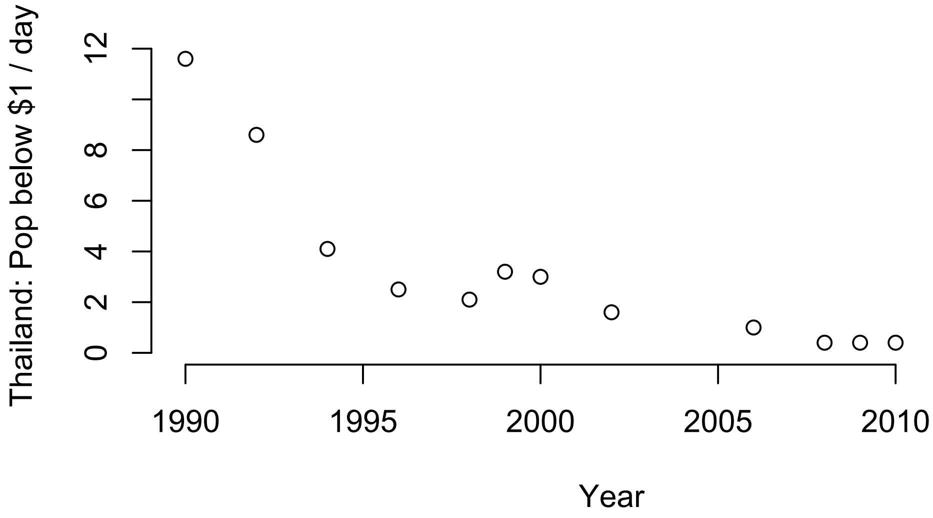

Table 5
Time series for Pct. Population under $1 /day for Thailand.

The change point procedure generates two change points – at 1994 and 2000 – and thus three segments. This yields three estimates for the mean in those segments: 8.1, 2.7, and 0.76 percent. The tests statistics for the change points are approximately Chi-squared and observed at 47.2 and 4.1 – both have p-values of less than .001.

The tests for change in variance, given these change points in mean, are approximately Chi-squared and observed at 3.2 and 2.08 for p-values of less than .05 and .10, respectively. The estimators for the variance are 14.25, 0.24, and 0.28.

There is strong evidence for change in mean across the three segments and change in variance across the first two segments.

Our test for trend is significant and downward across the series: there is a significant reduction in percent population under $1/ day in Thailand over 1990-2014. The series appears non-stationary from the first to second segments – i.e. from before and after 1994 – but stationary afterwards.

Our test for monotonicity "fails": the data appear to be monotonically decreasing across the length of the series and the rate of change of the trend– modulo the variation in each segment – is consistently negative.

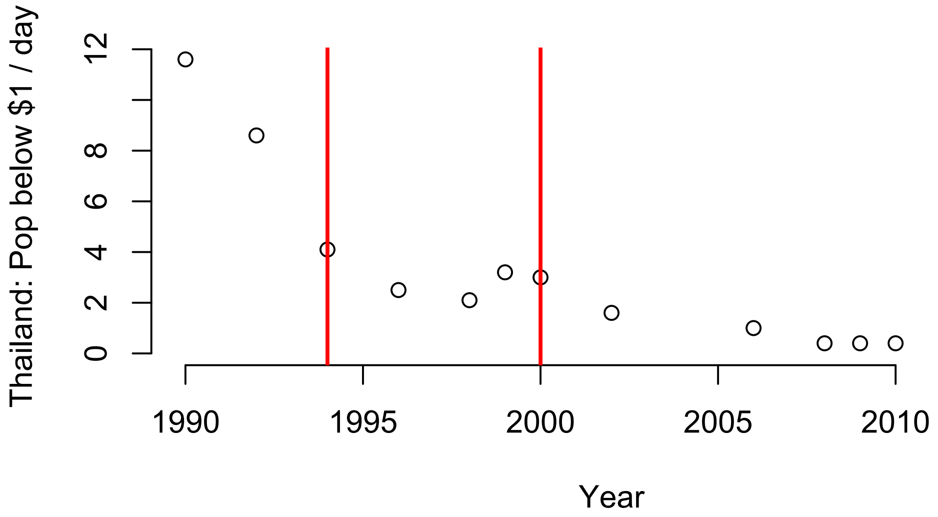

Table 6

Time series for Pct. Population under $1 /day for Thailand with segments identified by binary segmentation in vertical red lines.

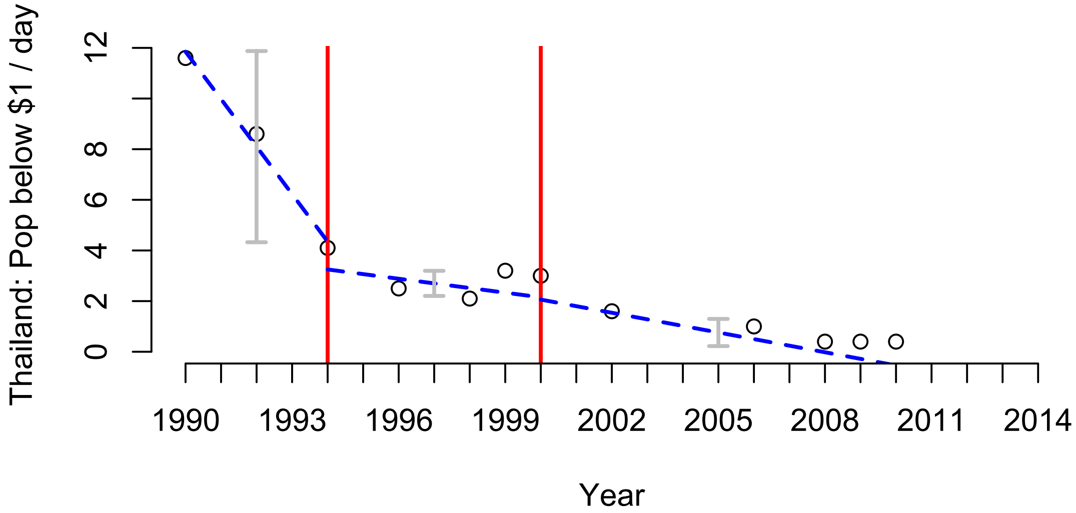

Table 7

Time series for Pct. Population under $1 /day for Thailand with segments, means/trend and variances (in red, blue and grey) identified. These estimators are immediately applicable to the univariate and multivariate random effects models.

To summarize: we look at four aspects of each data series as a first step and prelude to the univariate prediction of progress.

- Missingness – to assess the availability and possible resolution for the time series
- Trend – to determine change of the mean in the probability model
- Stationarity – to determine the variation in the probability model
- Monotonicity – to determine the rate of change in the parameters of the probability model.

These make sense in this order: some series are completely or almost completely missing which prevents prediction; trend is progress towards an MDG; stationarity affects measurement of that progress; and in a way, monotonicity captures both the validity of progress and/or its measurement.

Let's repeat this example with a different data series for a different country.

Example II

In this example we take the data series Percent Growth rate of GDP per person employed for Ecuador. Out of a possible 25 observations, 10 are available.

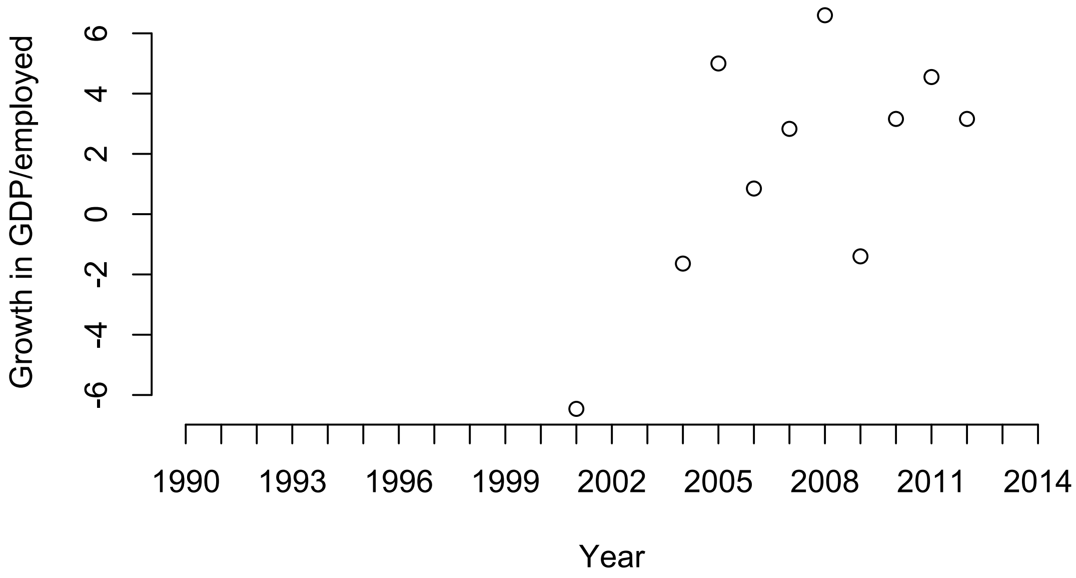

Table 8
Time series for Growth in GDP/employed for Ecuador.

The series is missing until 2001, then again until 2004, then again after 2012. The change point procedure generates two change points – at 2006 and 2009. This yields three estimates for the mean on these segments: -0.5625, 2.22 and 2.365 percent .

The test statistics for the change points are approximately Chi-squared and observed at 16.54 and 0.67 – the p-value for the first change point is less than .0001; the p-value for the second change point is above 10 percent.

Trend is significant – positive and increasing - across the first change point, then, but insignificant across the second change point.

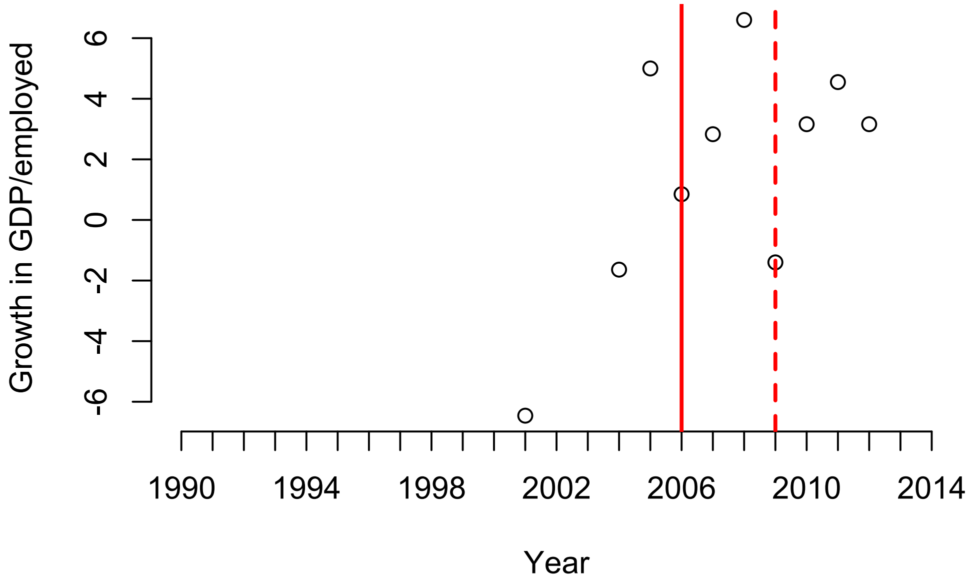

Table 8

Time series for Growth in GDP/employed for Ecuador with change point. Notice that the second change point (in 2009) is marked by a dashed line – the significance level for the test for this point is above ten percent.

The tests for change in variance, given these change points in mean, are approximately Chi-squared and observed at 1.22 and 0.22 for p-values of greater than .10 for both. The estimators for the variance are 22.95, 11.51, and 6.73.

The test for stationarity is insignificant: the variance appears stable (though large) across the change points.

The test for monotonicity is insignificant despite the appearance of the data: *the segment between the change points appears to be decreasing, yet the second change point is statistically insignificant.*

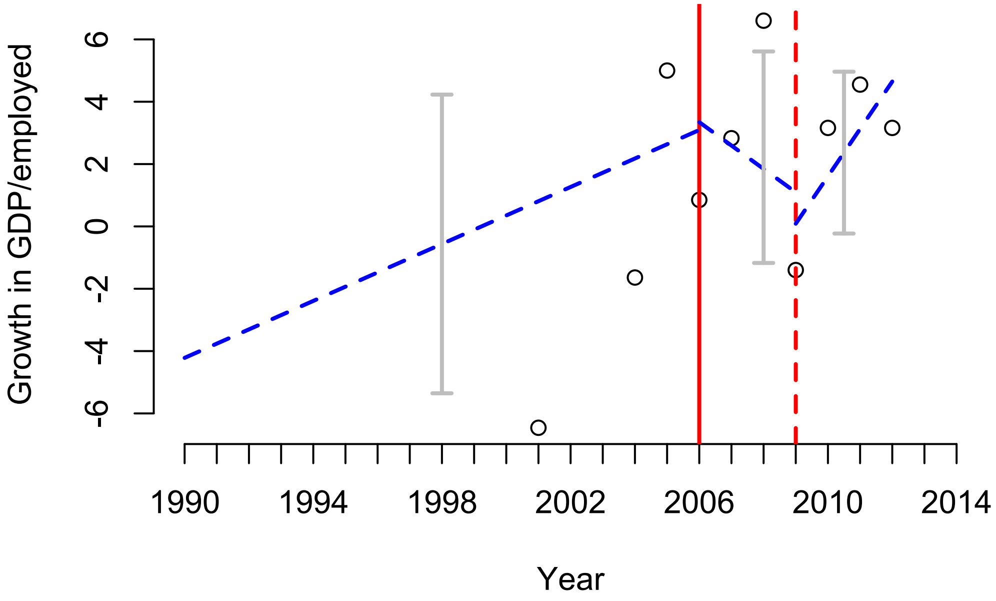

Table 9
Time series for Growth in GDP/employed for Ecuador, means, trend and variances (in red, blue and grey in order) identified. These estimators are immediately applicable to the univariate and multivariate random effects models. The dashed red line is the second change point – which is statistically insignificant. See Table 10.

Table 9 illustrates the importance of actively inspecting each data series before modeling. The change point analysis yields three segments however the second change point is only marginally significant. There is a clear increase in the values of series in the first segment; the picture in the second and third segments is not so clear.

Given the first change point at 2006 and ignoring the second change point at 2009: the series on the second segment has a downward trend with less variation *since the variation and mean are calculated across the second change point*.

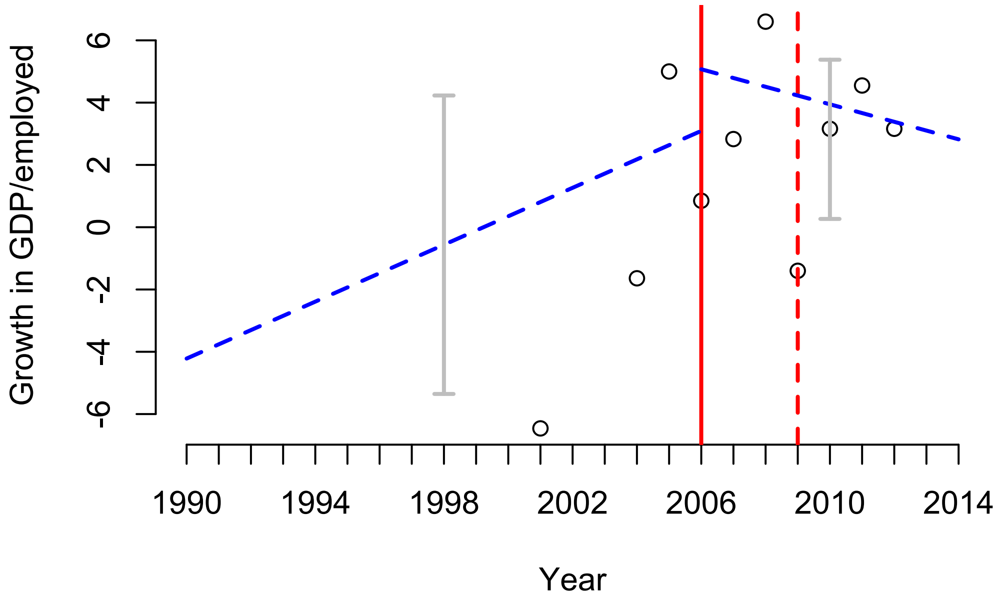

Table 10

Time series for Growth in GDP/employed for Ecuador, means, trend and variances (in red, blue and grey in order) identified. These estimators are immediately applicable to the univariate and multivariate random effects models. The dashed red line is the second change point – which is statistically insignificant. Here the change point at 2009 is considered an ordinary datum; one trend line is calculated from 2006 onward.

The contrast between Tables 9 and 10 is substantively a choice between beliefs about the real world phenomena the data illustrate. *Does the decrease in growth rate in 2009 in Ecuador represent a real change in the `on the ground' conditions?* Here is where exogenous knowledge – the very kind of information a specialist or analyst in the client country would have.

Choosing the model from Table 9 or Table 10 yields very different predicted values. In Table 9 the final trend line is positive – in Table 10 it is relatively flat but still negative. The 2015 prediction for Table 9 will be greater than that from Table 10.

Univariate-Based Prediction of MDG Progress

We use a linear random effects model with a discrete time Wiener process for error to generate the univariate predictions of MDG progress. These predictions are estimates of the observed values of the variable series in 2015 based upon the available data from 1990-2014.

We choose this model for several reasons:

- *Bayesian* updating: the parameters of this model (intercept and slope) are random in this framework and the model predictions incorporate not just measurement but modeling error.
- Discrete time process: The variable series for almost all the countries has irregularly spaced and highly missing data. Ordinary time series methods cannot be validly applied here. The discrete time process allows us to estimate the model parameters at the time points where the data are. We can propagate the Wiener process between irregularly spaced observations
- Piecewise-ness: The model is piecewise linear between observations – this is because the intercept and slope parameters can be updated at each observation.

Example I

Let's look at the series for Pct. Population under $1 /day for Thailand again as an example.

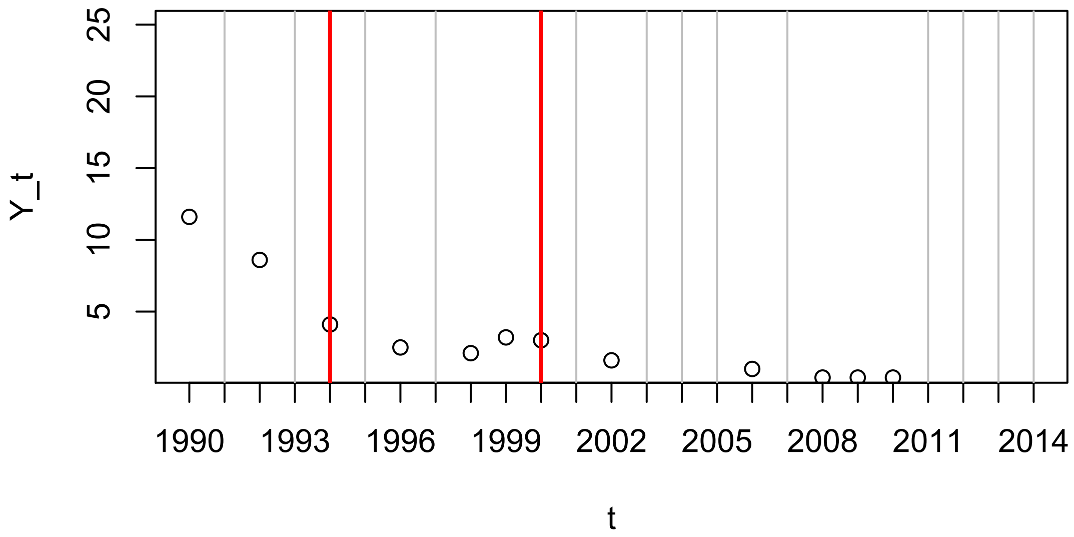

Table 11

Time series for Thailand, Pct. Population under $1/day. Observed values are in circles. There are irregularly spaced missing values, at: 1991, 1993, 1995, 1997, 2001, 2003-2005, 2006, 2011, 2014

Look at Table 8. The change point procedure yields two change points and thus three segments. Within each segment the mean and variance estimators from the change point procedure are plug in estimates for the `population wide' mean and variance of the series. The *hyperparameters* for the intercept and slope of the linear random effects model (see the Appendix) are generated from these estimators and are *posterior updated at each observed value*. In between observed values these hyperparameters change only with the time interval: *we do not use predicted values between the observed values to update the hyperparameters.*

The `population wide' mean and variance estimators are updated at each change point; thus the hyperparameters are reseeded.

In Table 8 the change points are highlighted with red vertical lines: at these years the hyperparameters are fully updated. The observed values are circles; *the vertical grey lines are the years where data are missing*. At the observed values the hyperparameters are updated by time step and the observed data; in between the observed values only the time step updates the hyperparameters.
The model yields a mechanism for generating random replicates at each year via the posterior distributions of the parameters. The mean value at each year can be used as the predicted value, given the model. What the model really offers is a

distribution for value at each year from which replicates can be generated, credible intervals and moments can be calculated.

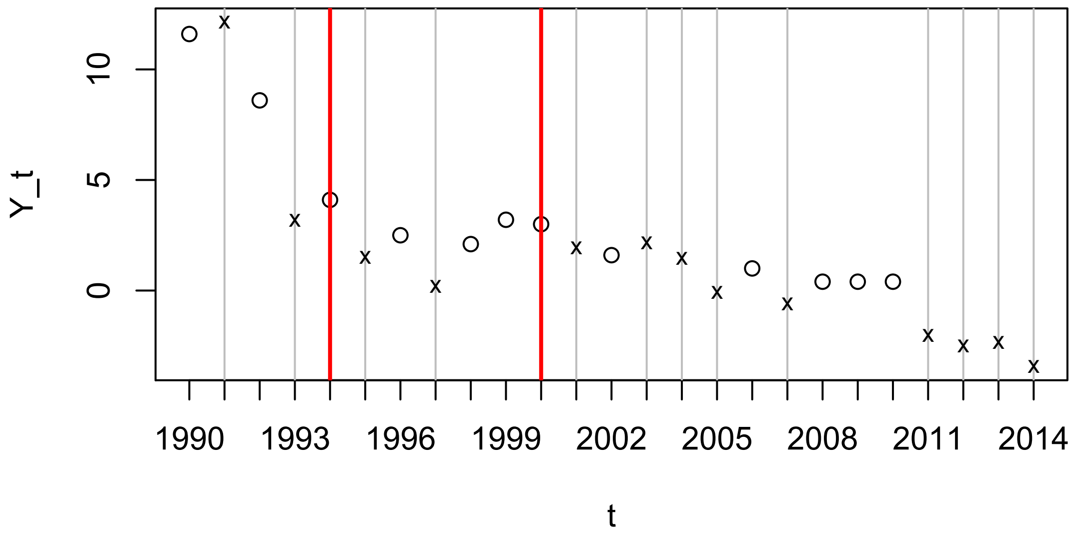

Table 12

Time series for Thailand, Pct. Population under $1/day. Observed values in circles, one particular set of predicted values in 'x'. Red vertical lines are the change points; grey vertical lines are years at which data are not observed. The y-axis on this graph is widened to illustrate the predicted values.

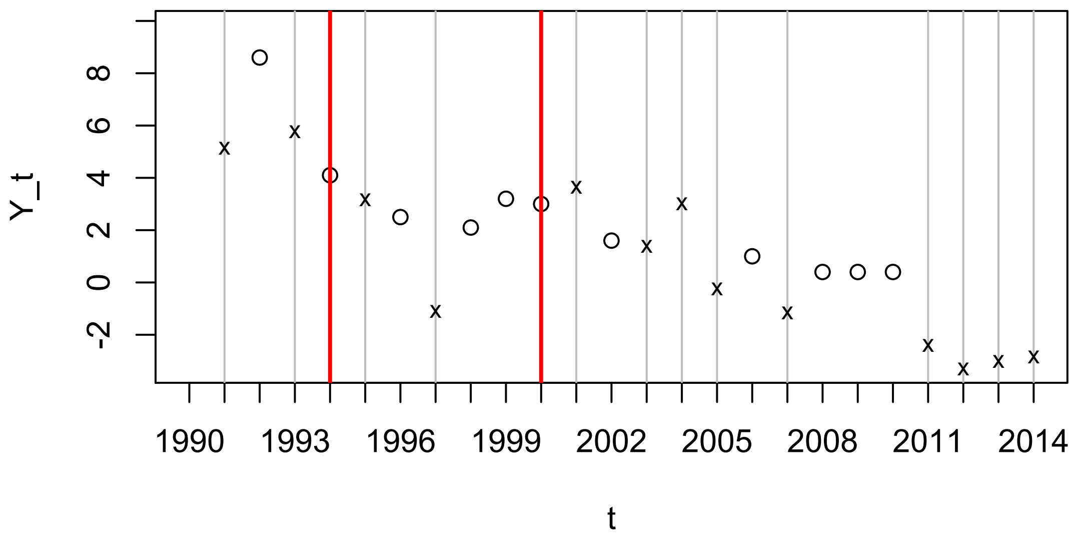

Table 13
Time series for Thailand, Pct. Population under $1/day. Observed values in circles, one particular set of predicted values in 'x'. Red vertical lines are the change points; grey vertical lines are years at which data are not observed. The y-axis on this graph is widened to illustrate the predicted values.

Tables 9 and 10 are illustrations of the observed values, change points, and two particular draws from the posterior distribution of the series. In most cases it is necessary to draw replicates to calculate the posterior mean and probability intervals: in this case the mean and variance of the posterior distribution are available in closed form

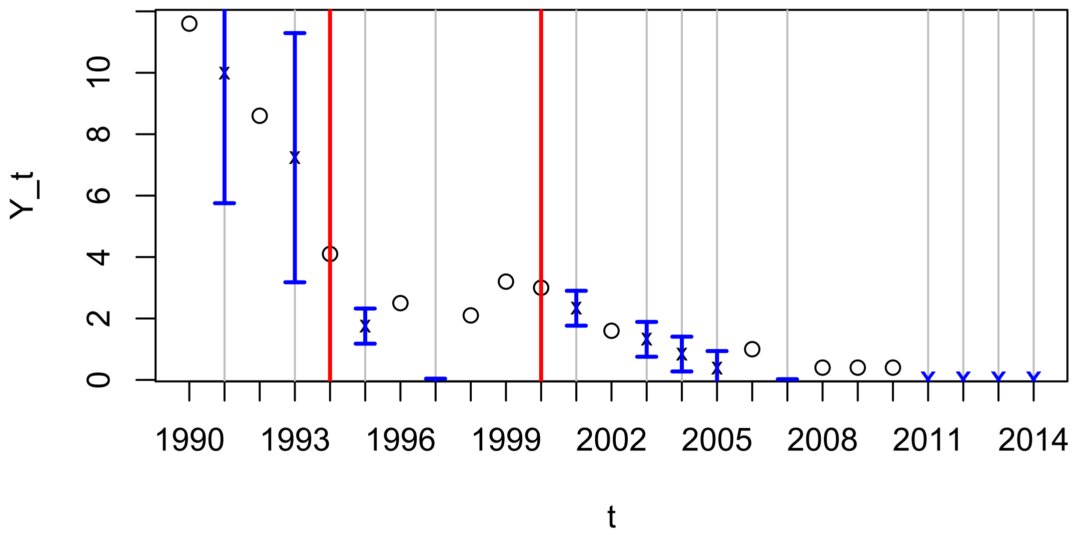

Table 14
Time series for Thailand, Pct. Population under $1/day. Observed values in circles; mean posterior values in black `x'; truncated predicted values in blue `x'.

Table 11 is an illustration of the predicted values from the model, with 95 % confidence intervals. Notice how the confidence intervals are quite large at the beginning of the series: there are few data available to estimate the posterior distribution and thus the variations in the predictions are large. After the first change point (in 1994) the variation decreases dramatically as the variance of the hyperparameters – given the change point – decreases and more data are available. At the end of the series the predicted values are truncated at zero.

In direct language the prediction for the percent of the population under $1/day in 2015 is zero percent. The model predicts a point estimate of -2.85 percent with a 95% credible interval of [-3.41, -2.29]. All of these values are below the practical floor of 0%.

These negative predictions are an artifact of the linear model: a three piece piecewise linear random effect model is fit to the entire data (see Table 7) given the two change points. The (random) slope of each of the segments is negative: the expectation of the posterior for the series is negative as well, then, at 2015. Any hypothesis test including positive values for this prediction in the alternative would fail. We fix the prediction at zero.

We adjust the model below to account for this limitation in piecewise linear random effects at the end of the time series.

Example II

Let's look at the change in GDP Growth rate for Ecuador again.

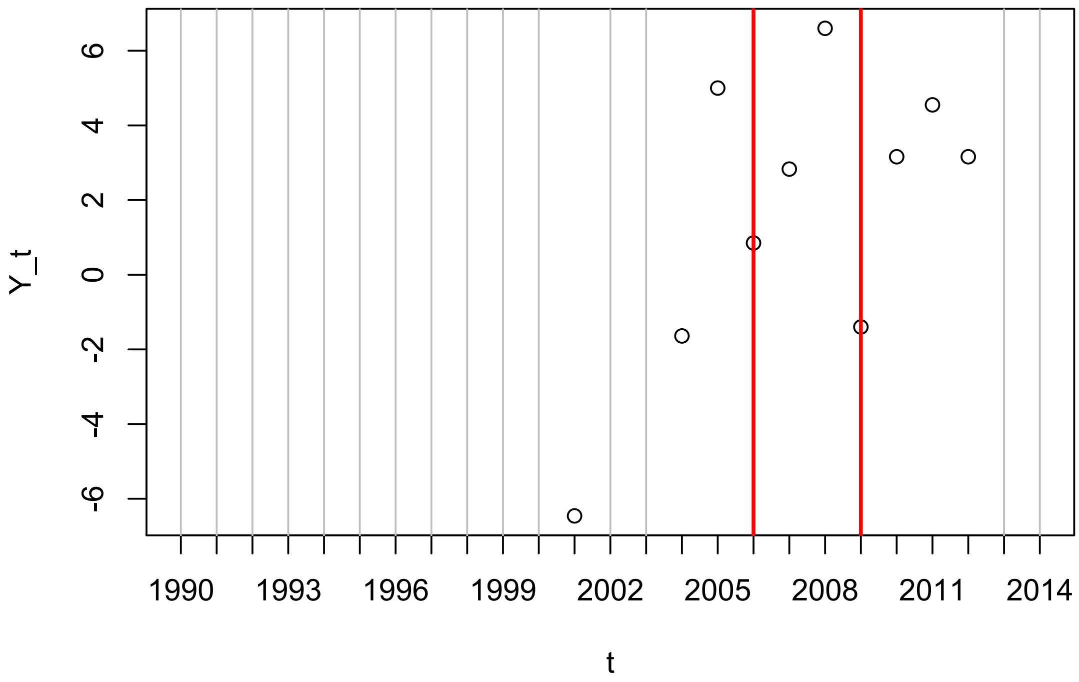

Table 15
Time series for Ecuador Change in GDP Growth Rate. Observed values are in circles. There are irregularly spaced missing values, at: 1991-1999, 2002-2003, 2013-2014

One way this example differs from Example I is that most of the missing values are at the beginning of the series. Another feature to note is the drop in growth rate in 2009 after what appears to be an upward trend. Notice as well that the change point procedure marks this observation.

We know – exogenously – that the measurement in 2009 reflects a worldwide drop in GDP/GDP growth rates and regard it as proper to in fact include this observation as an actual change point. This, of course, affects the model: we have three piecewise linear segments instead of two (Table 10 in favor of Table 9).

The estimators for the posterior distribution of the slope and intercept of the linear segments are sensitive to the change points and `new' observations (ordered in time). The inclusion of 2009 as a change point has a greater weight in the model and thus changes the predictions. We can see this in Table 16.

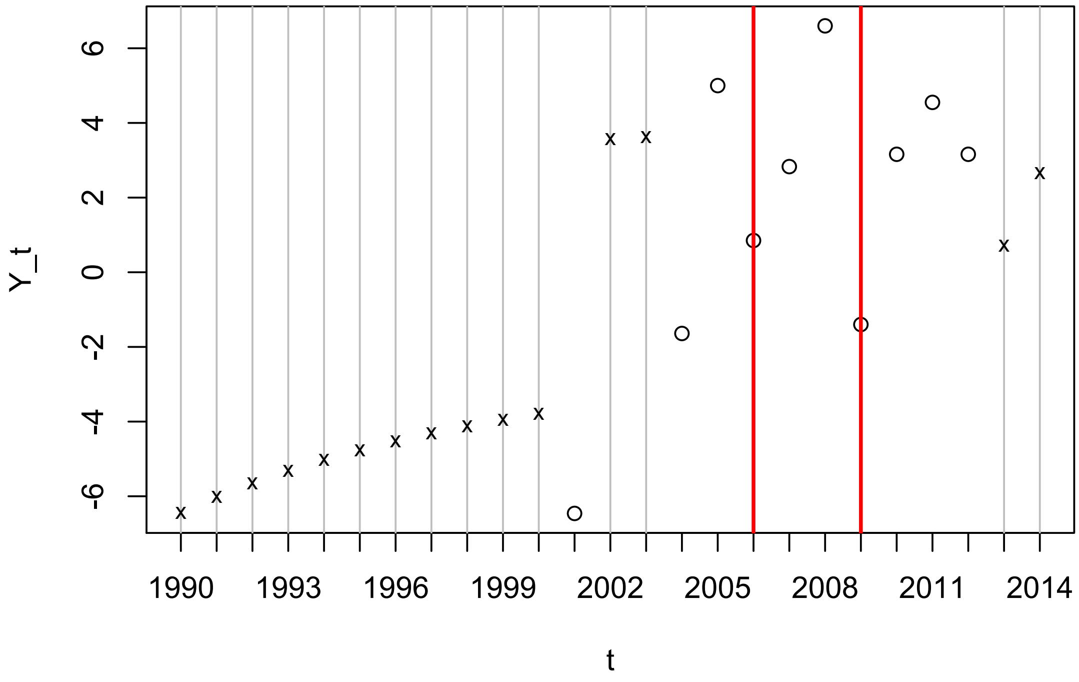

Table 16
Time series for Ecuador Change in GDP Growth Rate. Observed values are in circles, one particular set of predicted values in 'x'. Red vertical lines are the change points; grey vertical lines are years at which data are not observed.

The predicted values – again these are just one random draw from the posterior distribution – at the beginning of the series reflect this. These values are draws from the posterior of a linear model with an intercept with negative mean and slope with positive mean. As well the predicted values at the end of the series are lower than those from a model where 2009 is not a change point. This all makes sense and we point out how the modeling should well reflect the phenomena on the ground.

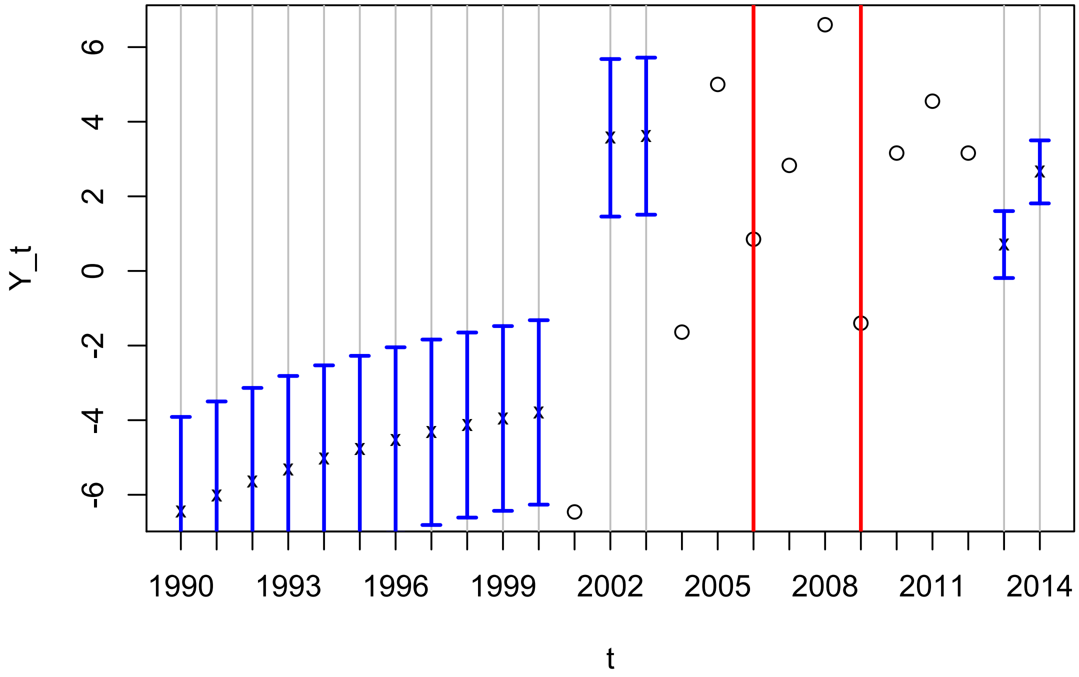

Table 16
Time series for Ecuador Change in GDP Growth Rate. Observed values in circles; mean posterior values in black `x'; posterior – credible – intervals for predicted values in blue error bars.

The error bars – these are from the standard deviation of the posterior distribution of the linear piece-wise model – at the beginning of the series are wider as there is no observed data to `fix' the model. The error bars for the last two predicted values (2013 and 2014) are much more narrow.

In direct language the univariate prediction for Ecuador for the Change in GDP Growth rate for 2015 is 4.92 percent with a 95% credible interval of [4.07, 5.76].

Multivariate-Based Prediction of MDG Progress

We augment the univariate based predictions by using additional data from the World Bank Development (WDI) indicators as predictors with *fixed effects* to each univariate model.[4]

Our approach is to model the differences between the predictions of the univariate model – we are using posterior means via a random effect model – via projection pursuit regression against a set of many covariates. *Projection Pursuit Regression* (PPR) consists of linear combinations of transformations of a set of explanatory variables. Here, we are balancing the interpretability of the linear random effects with the 'predictivity' of the PPR procedure – which as a standalone method suffers from interpretability.

Example I

We continue with the Thailand, Pct. Population under $1/day example. There are 13 missing values from 1990-2014; we fit the fixed effects model to the 12 available data. We restrict the WDI indicator set for Thailand to available data from 1990-2014.[5]

[4] http://data.worldbank.org/data-catalog/world-development-indicators

[5] Across the entire WDI data from 1990-2014 there are 57 percent missing values. We restrict predictors to WDI series that are completely observed over the time period.

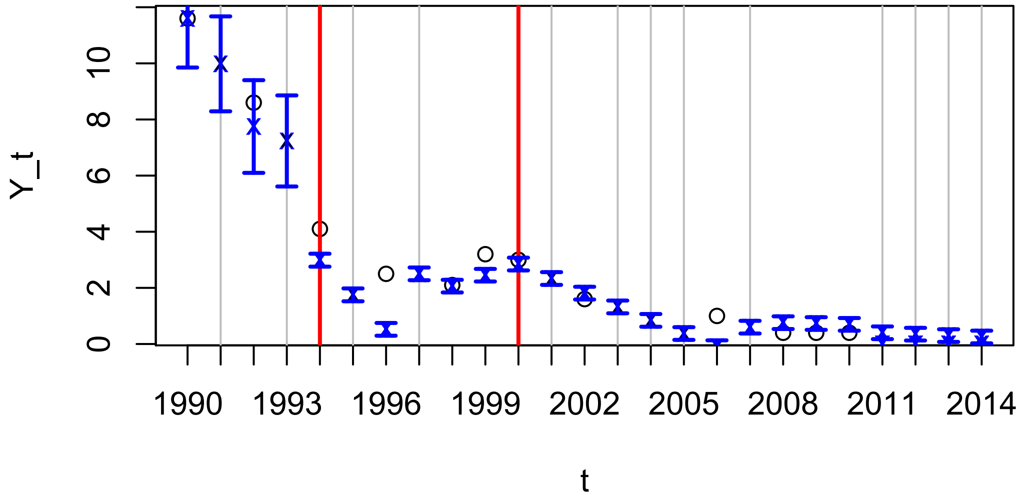

Table 17

Time series for Thailand, Pct. Population under $1/day. Observed values in 'o'; predicted values from mixed effect model in 'x'; 95% confidence intervals in blue; change points in red vertical lines; unobserved years in grey vertical lines.

Compare table 11 to table 10. Affixing the fixed effect PPR model to the univariate predictions narrows the confidence intervals[6] and yields predictions at the end of the series that are strictly positive and thus more reasonable.

The 2015 prediction for this series is a point estimate of 0.197 percent with a 95% confidence interval of [0.08,0.31].

[6] The mean squared error decreases by 40 percent for this example. The `posterior' confidence intervals from the univariate model are replaced with the model based variance from the mixed effect model.

Example II

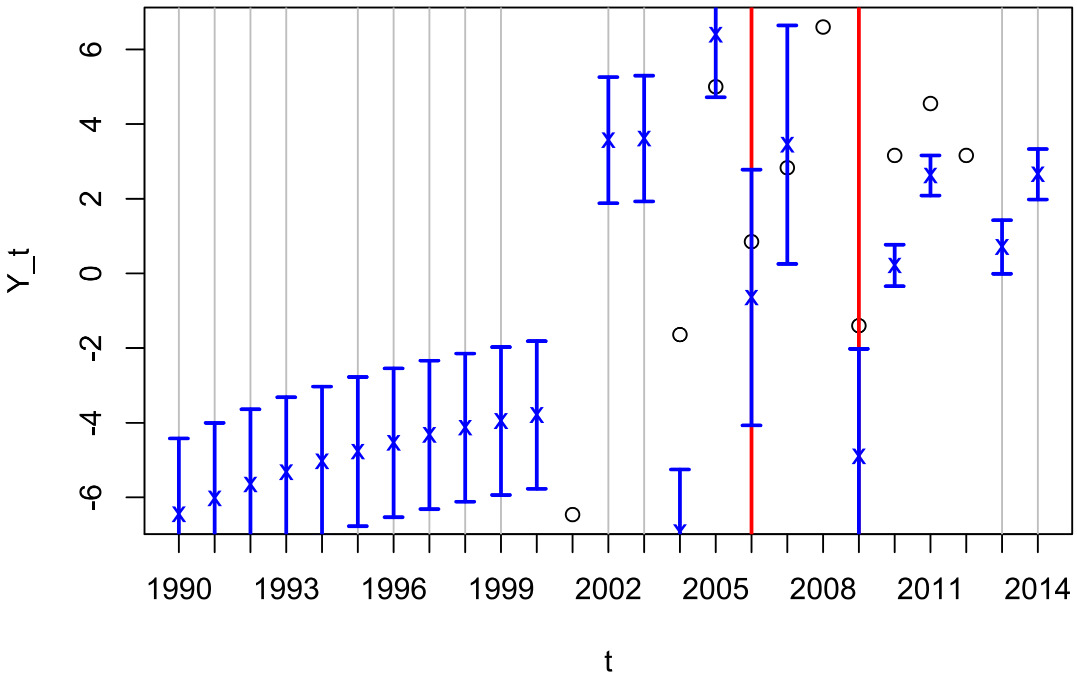

Table 18

Time series for Ecuador Change in GDP Growth Rate with predicted values from piecewise linear model augmented with Project Pursuit Regression (PPR). Observed values in 'o'; predicted values from mixed effect model in 'x'; 95% confidence intervals in blue; change points in red vertical lines; unobserved years in grey vertical lines.

Contrast Table 18 with Table 16: the PPR model has narrow confidence intervals at the beginning and end of the series. Notice that the confidence intervals are wider at the beginning of the series than at the end and especially wide at the observed values between the change points. The fixed effect PPR estimators are minimizing error across the entire series[7] – where the observed values are anomalous (i.e. at the change points) the predicted values from the (now) mixed effect model have more variance in account. Yet, the predicted values at the end of the series have narrower intervals; the model has the entirety of the data and the change points to lessen variation in the estimators (the random slope and intercept, and the error process).

The predicted value for the Growth Rate in GDP in Ecuador for 2015 is 5.26 percent with a 95% credible interval of [4.83 ,5.68]. This interval is wholly within the prediction interval of the random effects model and is far narrower.

[7] See the Appendix. In fact the PPR model is `fit' to the error process from the univariate random effects model.

MDG Data Analysis October 2014

Summary

This document is meant to be a guideline for the statisticians and researchers in the client countries for the MDGs to generate their own reports and predictions on their progress towards the MDG goals. With that in mind we have tried to balance sophisticated models against and clear and defensible assumptions.

In particular we suggest practitioners take our modeling procedures in order:

- We regard the missing data analysis, time series plots, tests for stationarity, trend and monotonicity as crucial first steps for modelers. In fact we believe the descriptive statistics and simple tests are zeroth order models. These sorts of analyses should be replicated – and should be easily replicated – by every client MDG country
- We chose piecewise linear models *after* change point procedures. The information from the change point procedures allows for conclusive tests of stationarity, trend and monotonicity. *A fortiori*: we regard the meaning and conclusion of these tests as very important guidance for the MDGs. Is the variable changing so rapidly that it is unpredictable?; are conditions on the ground getting better (or worse) rapidly?; are we losing or accelerating progress toward the goal? – these are the real questions these tests are designed to answer.
- The piece-wise linear models (given the change point) are not much more difficult to interpret than ordinary linear/least squares type models. But the random effects (random slope, random intercept) and time-stepped Bayesian updating allow the model to incorporate new data and flexibly reflect randomness in the observations
- The PPR modeling on the error process should be seen as exploitative of unknown association among the WDI and (among) MDG variables. Here, and necessarily only lastly, we have sacrificed interpretability for narrower error intervals about predictions.

The methods used here, while tailored to the specific data, should be replicable by reasonably sophisticated practitioners. The graphs and change point procedures can be produced in a spreadsheet or rudimentary statistics program. The random effects model (because of the use of the Gaussian posterior) has closed form estimators and does not require simulation or resampling. Lastly, the PPR regression is just a special case of matrix orthogonalization and can be computed easily as well.

Figures

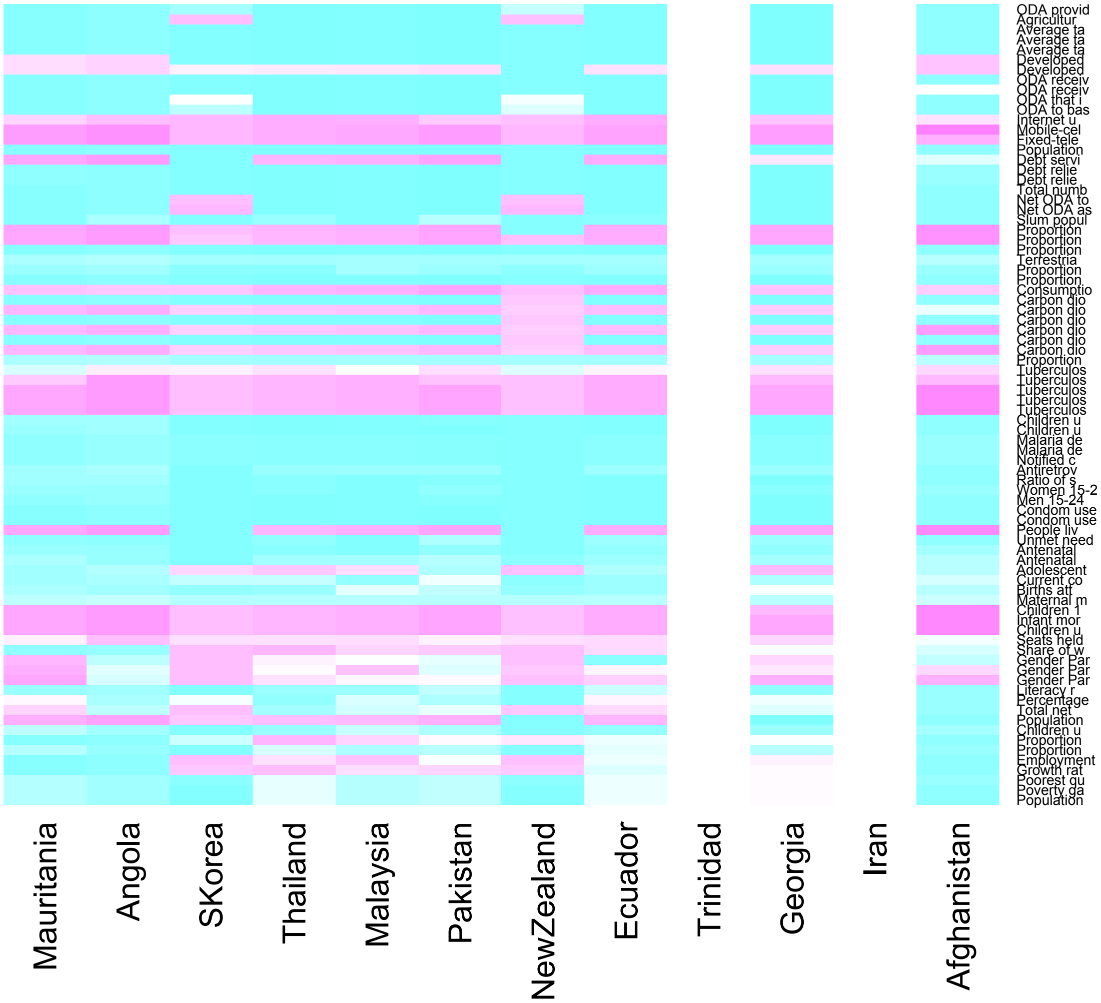

Figure 1

Figure 1 illustrates the missingness of the MDG data across 12 representative countries, over the 25 yearly observations from 1990-2014. Darker colors indicate less missingness, white – completely missing. Some countries have more missing values than others: notice both Trinidad and Iran are completely missing. Some data are less missing (across countries) than others – for example, measurements of Tuberculosis levels are well observed across all countries.

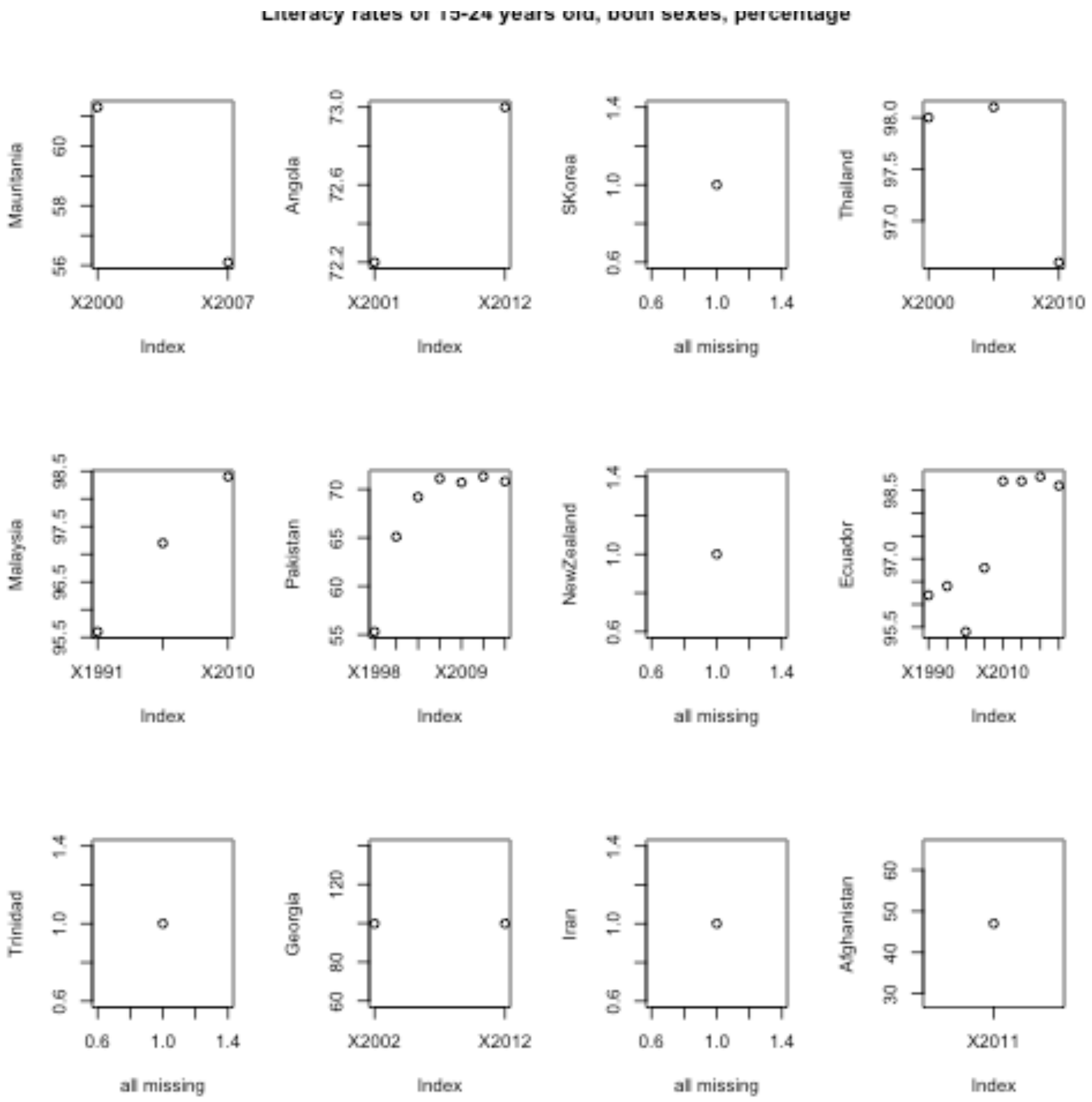

Figure 2

Illustration of data across countries for Percentage Literacy Rates of 15-24 year olds, both sexes. This variable is missing completely for South Korea, New Zealand, Iran, Trinidad and Afghanistan – removing the possibility of univariate prediction for those countries. For other countries – Thailand & Pakistan, for example – the data do not exist for all years. Univariate point predictions for these countries are less certain (have wider confidence intervals)

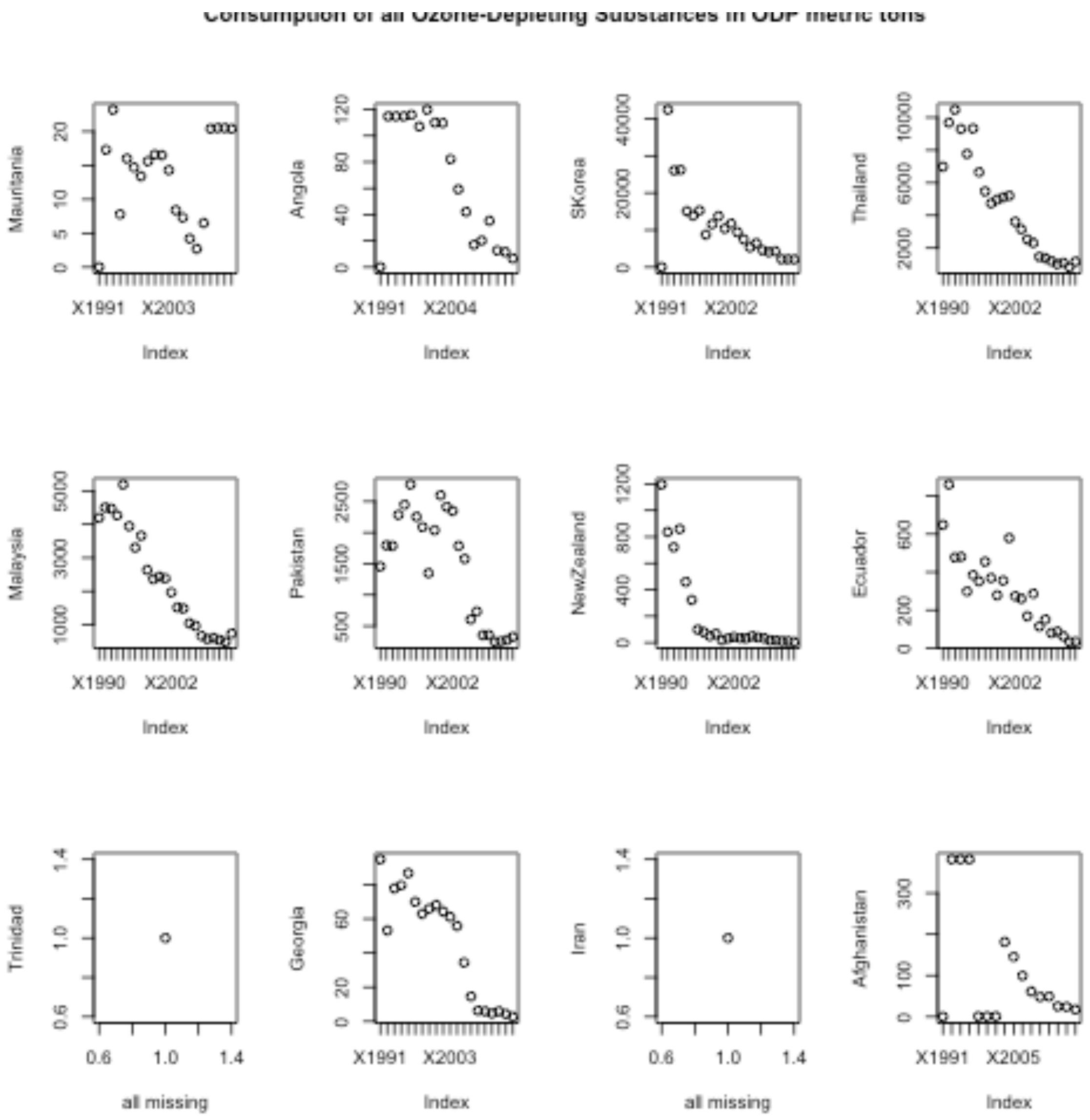

Figure 3

Illustration of data across countries for Consumption of an Ozone-Depleting Substance in Metric Tons. This series is well represented across most of the countries (save Trinidad and Iran).

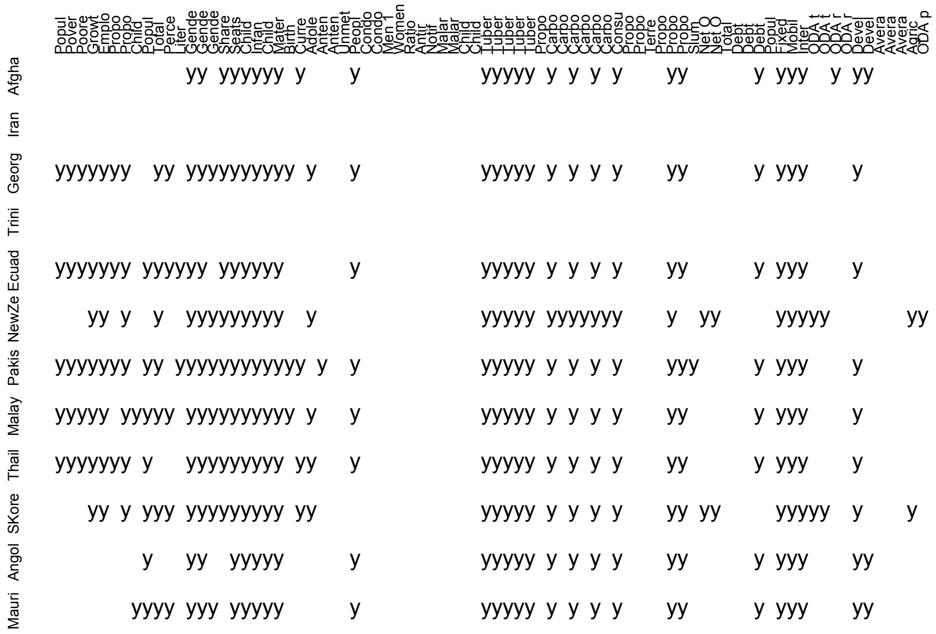

Figure 4

Diagram of MDG data used to predict MDG progress. Variable series with more than 80 percent missing values are not used for predictions. The available data is similar across countries.

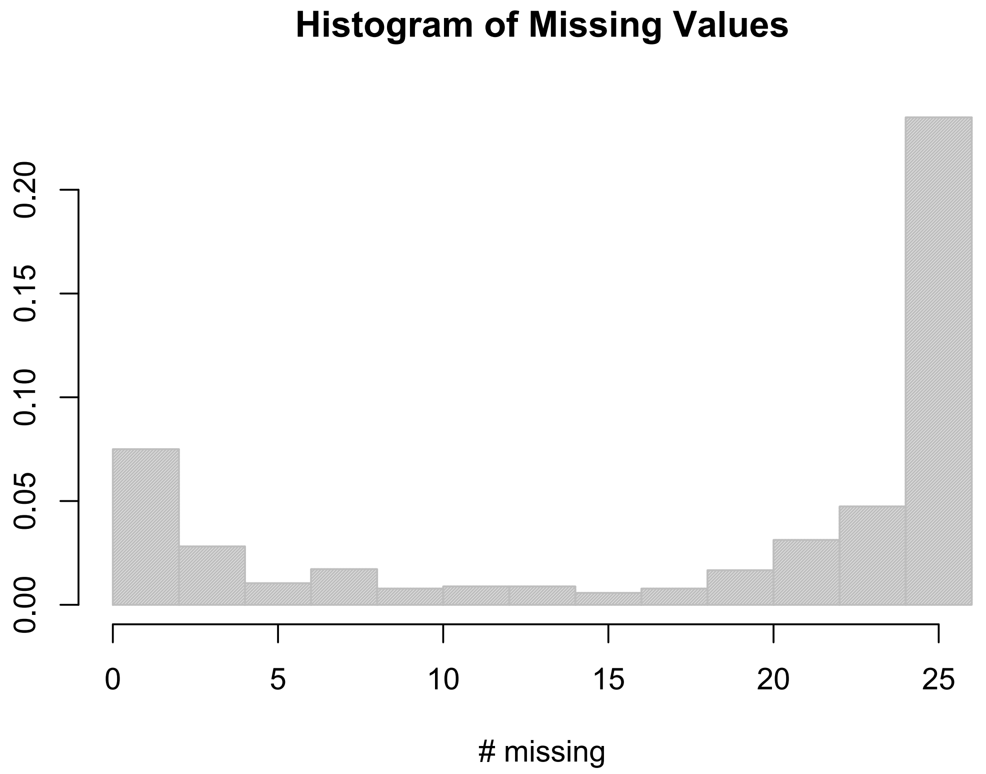

Figure 5

Histogram of missingness in MDG data. The observed counts are by variables (80 variables) Many variables – more than a fifth of the variables in the selected countries – are completely missing.

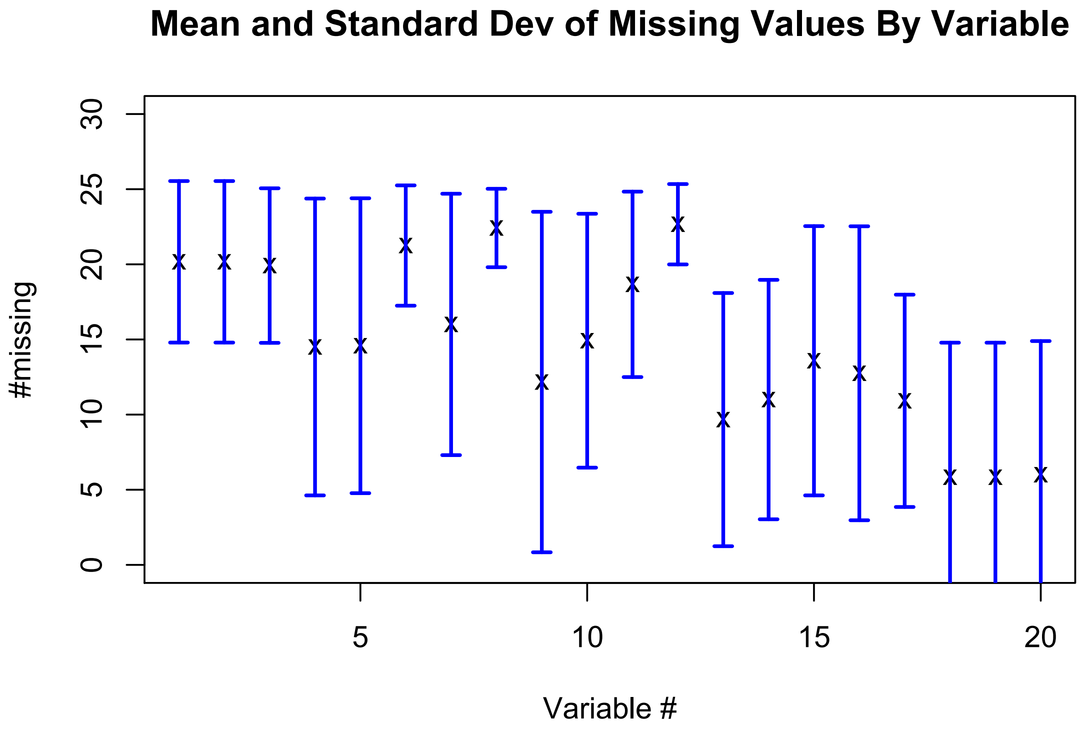

Figure 6
Plot of missingness in MDG data by variable with 1 standard deviation in blue. Variables 1-20. See Figure 10 for list of MDG variables.

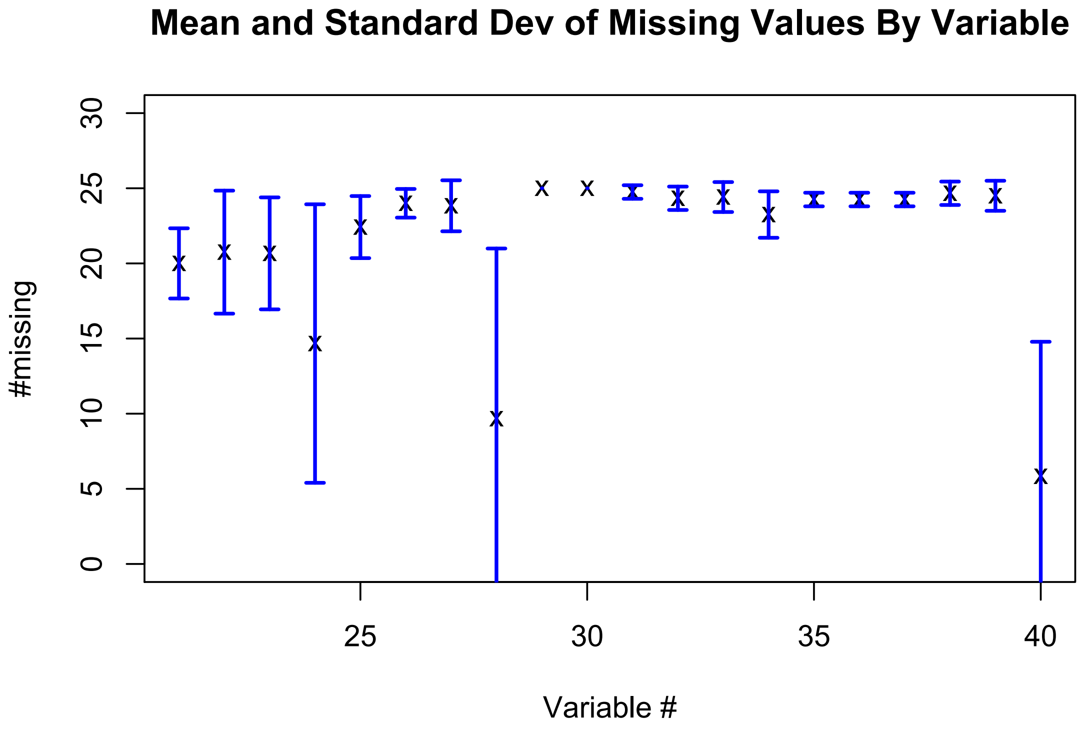

Figure 7

Plot of missingness in MDG data by variable with 1 standard deviation in blue. Variables 21-40. See Figure 10 for list of MDG variables.

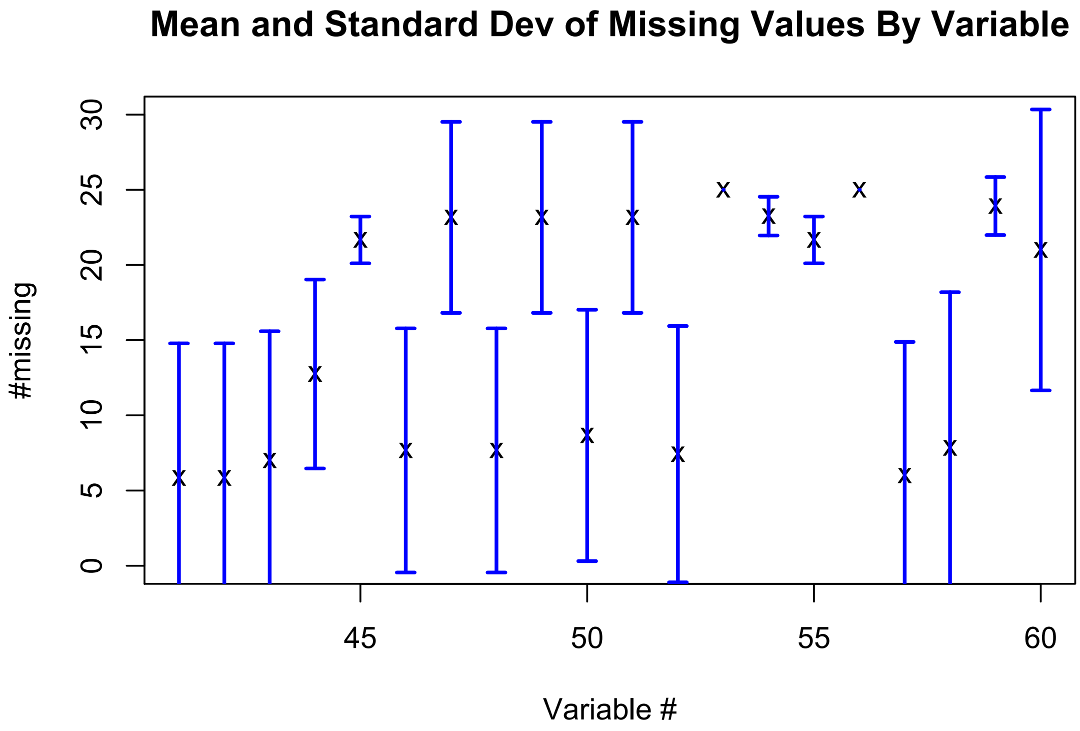

Figure 8

Plot of missingness in MDG data by variable with 1 standard deviation in blue. Variables 41-60. See Figure 10 for list of MDG variables.

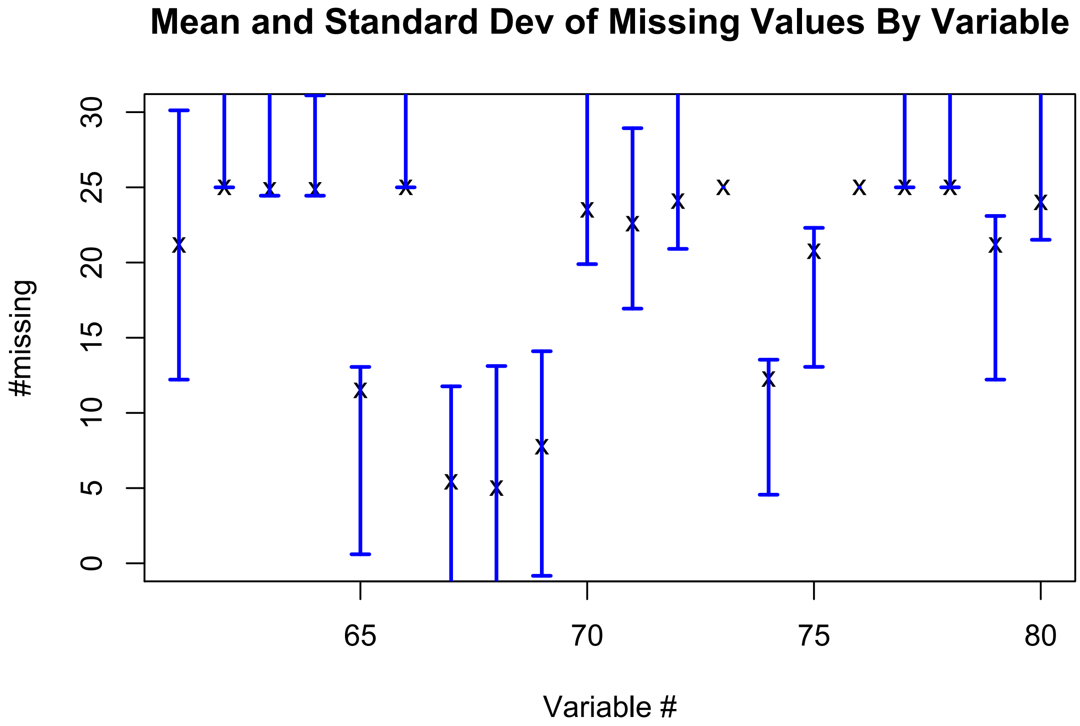

Figure 9

Plot of missingness in MDG data by variable with 1 standard deviation in blue. Variables 61-80. See Figure 10 for list of MDG variables.

MDG Data Analysis   October 2014

1 Population below $1 (PPP) per day, percentage

2 Poverty gap ratio at $1 a day (PPP), percentage

3 Poorest quintile's share in national income or consumption, percentage

4 Growth rate of GDP per person employed, percentage

5 Employment-to-population ratio, both sexes, percentage

6 Proportion of employed people living below $1 (PPP) per day, percentage

7 Proportion of own-account and contributing family workers in total employment, both sexes, percentage

8 Children under 5 moderately or severely underweight, percentage

9 Population undernourished, percentage

10 Total net enrolment ratio in primary education, both sexes

11 Percentage of pupils starting grade 1 who reach last grade of primary, both sexes

12 Literacy rates of 15-24 years old, both sexes, percentage

13 Gender Parity Index in primary level enrolment

14 Gender Parity Index in secondary level enrolment

15 Gender Parity Index in tertiary level enrolment

16 Share of women in wage employment in the non-agricultural sector

17 Seats held by women in national parliament, percentage

18 Children under five mortality rate per 1,000 live births

19 Infant mortality rate (0-1 year) per 1,000 live births

20 Children 1 year old immunized against measles, percentage

21 Maternal mortality ratio per 100,000 live births

22 Births attended by skilled health personnel, percentage

23 Current contraceptive use among married women 15-49 years old, any method, percentage

24 Adolescent birth rate, per 1,000 women

25 Antenatal care coverage, at least one visit, percentage

26 Antenatal care coverage, at least four visits, percentage

27 Unmet need for family planning, total, percentage

28 People living with HIV, 15-49 years old, percentage

29 Condom use at last high-risk sex, 15-24 years old, women, percentage

30 Condom use at last high-risk sex, 15-24 years old, men, percentage

31 Men 15-24 years old with comprehensive correct knowledge of HIV/AIDS, percentage

32 Women 15-24 years old with comprehensive correct knowledge of HIV/AIDS, percentage

33 Ratio of school attendance rate of orphans to school attendance rate of non orphans

34 Antiretroviral therapy coverage among people with advanced HIV infection, percentage

35 Notified cases of malaria per 100,000 population

36 Malaria death rate per 100,000 population, all ages

37 Malaria death rate per 100,000 population, ages 0-4
38 Children under 5 sleeping under insecticide-treated bed nets, percentage
39 Children under 5 with fever being treated with anti-malarial drugs, percentage
40 Tuberculosis prevalence rate per 100,000 population (mid-point)
41 Tuberculosis death rate per year per 100,000 population (mid-point)
42 Tuberculosis incidence rate per year per 100,000 population (mid-point)
43 Tuberculosis detection rate under DOTS, percentage (mid-point)
44 Tuberculosis treatment success rate under DOTS, percentage
45 Proportion of land area covered by forest, percentage
46 Carbon dioxide emissions (CO2), thousand metric tons of CO2 (CDIAC)
47 Carbon dioxide emissions (CO2), thousand metric tons of CO2 (UNFCCC)
48 Carbon dioxide emissions (CO2), metric tons of CO2 per capita (CDIAC)
49 Carbon dioxide emissions (CO2), metric tons of CO2 per capita (UNFCCC)
50 Carbon dioxide emissions (CO2), kg CO2 per $1 GDP (PPP) (CDIAC)
51 Carbon dioxide emissions (CO2), kg CO2 per $1 GDP (PPP) (UNFCCC)
52 Consumption of all Ozone-Depleting Substances in ODP metric tons
53 Proportion of fish stocks within safe biological limits
54 Proportion of total water resources used, percentage
55 Terrestrial and marine areas protected to total territorial area, percentage
56 Proportion of species threatened with extinction
57 Proportion of the population using improved drinking water sources, total
58 Proportion of the population using improved sanitation facilities, total
59 Slum population as percentage of urban, percentage
60 Net ODA as percentage of OECD/DAC donors GNI
61 Net ODA to LDCs as percentage of OECD/DAC donors GNI
62 Total number of countries that have reached their HIPC decision points and number that have reached their HIPC completion points (cumulative)
63 Debt relief committed under HIPC initiative, cumulative million US$ in end-2009 NPV terms
64 Debt relief delivered in full under MDRI initiative, cumulative million US$ in end-2009 NPV terms
65 Debt service as percentage of exports of goods and services and net income
66 Population with access to essential drugs, percentage
67 Fixed-telephone subscriptions per 100 inhabitants
68 Mobile-cellular subscriptions per 100 inhabitants
69 Internet users per 100 inhabitants
70 ODA to basic social services as percentage of sector-allocable ODA
71 ODA that is untied, percentage
72 ODA received in landlocked developing countries as percentage of their GNI
73 ODA received in small islands developing States as percentage of their GNI
74 Developed country imports from developing countries, admitted duty free, percentage
75 Developed country imports from the LDCs, admitted duty free, percentage

76 Average tariffs imposed by developed countries on agricultural products from developing countries

77 Average tariffs imposed by developed countries on textiles from developing countries

78 Average tariffs imposed by developed countries on clothings from developing countries

79 Agriculture support estimate for OECD countries as percentage of their GDP

80 ODA provided to help build trade capacity, percentage

Figure 10

Ordered list of MDG variables.

Technical Appendix

Data Quality Assessment

Let's call

$$Y_t$$

a random variable for an observed MDG data series, y_t. Associated data and random variables – used to assist prediction – are vectors $\boldsymbol{x}_t, \boldsymbol{X}_t$.

Here t is discrete on $(1990, 1991, \dots, 2014)$.

In all of the modeling here we make a standard assumption that the data (and parameters, and error) are generated by the Gaussian model. However, we choose modeling procedures that are flexible enough so that the Gaussian distributions are replaceable in a revision.

In the univariate modeling Y_t is is a discrete time linear process with a random intercept and a random slope

$$Y_t = \theta_0 + \theta_1 t + \varepsilon$$

where $\theta_0, \ \theta_1$ are independent parameters with *prior distributions* and ε is a discrete time martingale (expectation independent given history) process.

In terms of the MDG data we let *i* index across countries; *j* across variables/series; *t* is the discrete time index. So, the dimension of the data array, $y_{ij,t}$, is 80 x 81 x 25.

In order: we consider if there are enough data to model, if there is evidence of a change in mean, a change in variance, and stability in distribution.

We make no distributional assumptions in our tests for missingness; our tests for change in mean and variance (trend and stationarity) assume the Gaussian model but only through the likelihood ratio statistic; the tests for monotonicity are also derived via the likelihood ratio. The modeling in the univariate and multivariate sections use random effects; we could impose different priors and generate alternate posterior distributions algorithmically.

MDG Data Analysis   October 2014

Missingness

Let Z_{ij} be a random variable such that

$$\boldsymbol{P}\left(Z_{ij} = 1\right) \propto \boldsymbol{y}$$

and

$$z_{ij} = 1 \; when \; y_{ij} \; is \; missing,$$

where $\boldsymbol{y}$ are the observed data, then we call the data **y** *missing at random* (MAR), because the probability of unobserved data is not independent of the data. If

$$\boldsymbol{P}\left(Z_{ij} = 1\right) \perp \boldsymbol{y}$$

then we call the data *missing completely at random* (MCAR) as missingness is independent of the value of the data. If

$$\boldsymbol{P}\left(Z_{ij} = 1\right) = F_y(\boldsymbol{y})$$

where $F_y(\boldsymbol{y})$ is a probability distribution for the data, then the data are called *not missing at random* (NMAR) because the probability of being unobserved is completely dependent upon the value of the data.

Consider Figure 1, if the MDG data were MCAR we would expect to see a random distribution of shades in the cells of the heatmap matrix. Conversely, if the data were NMAR the missing values would be completely dependent upon the value and sort of data; we would expect to see a `perfect' pattern in the shades of the heatmap. Instead, we see several rows of almost complete missingness, indicating the variable itself is missing and independent of the particular observed data value.

For example the data on tuberculosis incidence appear to be MCAR. In other cases, for example missingness in the data on population and poverty (the lowest rows in the heat map) there appears to be some pattern in the missingness across countries and thus MAR.

We choose to only proceed with modeling when percent missing is below 80 percent. So we consider the estimator

$$\widehat{p_{\imath\jmath}} = \frac{num\ missing}{25} = \boldsymbol{P}(\widehat{z_{\imath\jmath} =} 1)$$

for the probability of missingness at each country-series. Next,

$$\widehat{p_{.\jmath}} = \sum_{i=1}^{I} \frac{p_{ij}}{I} = \boldsymbol{P}(\widehat{z_{.\jmath} =} 1)$$

and

$$\widehat{p_{\iota.}} = \sum_{j=1}^{J} \frac{p_{ij}}{J} = \boldsymbol{P}(\widehat{z_{\iota.} = 1})$$

are the estimators for the probability of missingness for a variable (across countries) and for a country (across variables): I and J are the number of countries and variables.

Trend

The hypothesis test in the change point procedure is:

$$H_o: \mu_t = \mu, \forall\ t$$
$$\text{vs.}$$
$$H_a: \exists\ t\ s.t. \mu_t \neq \mu$$

where t is some interval. Since the series are discrete, the t intervals are integer multiples of a year.

We assume the Y_t follow a Gaussian distribution with a constant mean in the null hypothesis; the alternative is that the mean changes.

The first step in our *modified binary segmentation* procedure is to partition the data, estimate the mean for both partitions, and use this partitioned likelihood as the first alternative.

Let

$$\Lambda(\mu, \sigma)$$

be the Gaussian likelihood for the entire available data series with μ, σ the values of the mean and variance under the broadest null hypothesis: no change points; no change in mean in variance.

The first step is to partition the data, let

$$\Lambda_1(\mu_L, \mu_R; \sigma)$$

be the Gaussian likelihood with μ_l, μ_r the plug in estimators for the mean to the left and the right of the first change point. This change point is identified exhaustively: Λ_1 is the maximum of the T possible, where T is the number of observations for the variable series.

Then

$$\frac{\Lambda_1(\mu_L, \mu_R; \sigma)}{\Lambda(\mu, \sigma)} \sim \chi^2(T-2)$$

is an approximately Chi-squared test statistic for the first change point.

The next step is a further partition, let

$$\Lambda_1(\mu_L\colon \mu_L^1, \mu_L^2; \mu_R\colon \mu_R^1, \mu_{R;}^2\sigma)$$

be the Gaussian likelihood after a repeat of the first step – exhaustive searches on each of the left and right partitions. Then

$$\frac{\Lambda_1(\mu_L\colon \mu_L^1, \mu_L^2; \mu_R\colon \mu_R^1, \mu_{R;}^2\sigma)}{\Lambda_1(\mu_L, \mu_R; \sigma)} \sim \chi^2(T-3)$$

is an approximately Chi-squared test statistic for the second and third change points.

Stopping here generates a maximum of three change points and four segments – thus four estimates for the mean. For these data, *T* is at minimum 5 and on average 7. Further segmentation is unreasonable for these data; for many series only one partition is possible.

Stationarity

For tests of stationarity we begin with

$$\Lambda(\mu, \sigma)$$

as the full Gaussian likelihood. The alternative likelihood, on the first partition, is

$$\frac{\Lambda_1(\mu_L, \mu_R; \sigma_L^2, \sigma_R^2)}{\Lambda(\mu, \sigma)} \sim \chi^2(T-3)$$

where we use the left and right means as plug in estimators. The second partition generates the following test statistic

$$\frac{\Lambda_1(\mu_L^1, \mu_L^2; \mu_R^1, \mu_{R;}^2\sigma_L^2, \sigma_{L'}^2, \sigma_R^2, \sigma_{R'}^2)}{\Lambda_1(\mu_L, \mu_R; \sigma_L^2, \sigma_R^2)} \sim \chi^2(T-5)$$

which is approximately Chi-squared distributed test statistic for the second and third change points.

Monotonicity

Let m_s be the slope for segment $s = 1, \dots, S$ with S number of segments defined by the change point procedure. Given the change points and segments (i.e. given the estimates of the mean and variance for each segment) the elements in the vector are *the* estimators for rate of change in the series in a piecewise linear model.

Our test for monotonicity in a series is essentially a test that the vector $\boldsymbol{m}$ resides in the first or third multivariate orthant.

Univariate Progress Predictions

The model we use for univariate prediction is linear, with a random slope and random intercept with discrete Brownian motion for error.

$$Y_t = \theta_0 + \theta_1 t + \varepsilon_t$$

where $\theta_0, \theta_1$are the random slope and intercept, distributed as

$$\theta_0 \sim N(\mu_0, \sigma_0^2)$$

and

$$\theta_1 \sim N(\mu_1, \sigma_1^2)$$

and

$$\varepsilon_t \sim N(0, \sigma^2 t)$$

their *prior* distributions.

This is a *Bayesian* approach that generates these *posterior* distributions for the parameters:

$$\theta_0 \sim N(\mu_{\theta_0}, \sigma_{\theta_0}^2)$$

and

$$\theta_1 \sim N(\mu_{\theta_1}, \sigma_{\theta_1}^2)$$

This approach allows the baseline expected value and rate of change of each variable to be updated as the series progresses. This model is piecewise linear, then, where each linear section can be updated at each new series observation.

Let time `epochs' be $t_1, \dots, t_k$where $t_j = j \cdot t_1, j = 1, \dots k$ with the corresponding series $Y_1, \dots Y_k$ with $1 \leq k \leq T$. Given the observed series $Y_1, \dots Y_k$ we can calculate, in closed form, the parameters of the posterior distribution of the intercept and slope at each k.

$$\mu_{\theta_0}(k) = A^{-1}[(y_1\sigma_0^2 + \mu_0\sigma^2)(\sigma_1^2 t_k + \sigma^2) - \sigma_0^2 t_1(\sigma_1^2 \sum_{t=1}^{k} y_t + \mu_1\sigma^2)]$$

$$\mu_{\theta_1}(k) = A^{-1}[(\sigma_1^2 \sum_{t=1}^{k} y_t + \mu_1\sigma^2)(\sigma^2 t_1 + \sigma_0^2) - \sigma_1^2(\sigma_0^2 y_1 + \mu_0\sigma^2 t_1)]$$

$$\sigma_{\theta_0}^2(k) = A^{-1}[\sigma^2\sigma_0^2 t_1(\sigma_1^2 t_k + \sigma^2)]$$

$$\sigma_{\theta_1}^2(k) = A^{-1}[\sigma^2\sigma_1^2(\sigma_0^2 + \sigma^2 t_1)]$$

with

$$A(k) = (\sigma_0^2 + \sigma^2 t_1)(\sigma^2 + \sigma_1^2 t_k) - \sigma_0^2\sigma_1^2 t_1$$

Here k is the index across years of the series and reset at each change point. This is a modeling choice: many of the series seem to vary wildly across change points; k could merely increase to incorporate each of the preceding values as well.

This generates T predicted values for each series as the mean of the posterior distribution of Y_t. The predicted values are not used where observed values are available nor are they used in the calculation of the posterior distribution. *Only the observed values affect the estimators of the posterior distribution.*

This approach is fully Bayesian and we do not propagate randomness further than the hyperparameters: we don't regard the hyperparameters of either the prior or posterior distributions as having any further randomness.

We take the predicted value for a series as the expectation of the posterior:

$$E(Y_t) = \mu_{\theta_0}(k = t) + \ t \cdot \mu_{\theta_1}(k = t)$$

as this is available in closed form for these priors its not necessary to generate replicates . Since the hyperparameters have no (i.e. a degenerate) distribution, the variance of the prediction is simply:

$$Var(Y_t) = \sigma_{\theta_0}^2(k = t) + t^2\sigma_{\theta_1}^2 - \frac{\sigma^2 t}{2}$$

Again we 'reset' k at each change point – we consider the change points as fixed and given with respect to the estimation of the linear models. Without change points the estimators for the hyperparameters would be based on the entire data series which would make the imposition of the linear model more constraining.

Multivariate Progress Predictions

We augment the univariate model by considering the *residuals* (the observed values of the discrete time process) as generated by a function – *via fixed effects* – on a multivariate set of data.

Let $\boldsymbol{x}_{\boldsymbol{\tau}}$ be a matrix of data for a particular country (not indexed), and τ the index set for the years of the univariate series. Partition $\boldsymbol{x}_\tau$ into:

$$\boldsymbol{x}_{\boldsymbol{\tau}} = (\boldsymbol{x}_{\tau_{m},}, \boldsymbol{x}_{\tau_o})$$

by indexing the missing and observed years of values of the univariate series y_t.

Let

$$e_t = Y_t - y_t = \theta_0 + \theta_1 t - y_t$$

be the observed value of the discrete time (Wiener) error process. This is a random variable with a posterior distribution of

$$e_t {\sim} N(\mu_{\theta_0} + \mu_{\theta_1} t(k), \sigma_{\theta_0}^2 + t(k)^2\sigma_{\theta_1}^2 - \frac{\sigma^2 t}{2})$$

We take this random variable – just the error process when we use the mean of the posterior as the predicted value – as the response to a *Projection Pursuit Regression* model on available data. The model

$$e_t = \beta_0 + \sum_{s=1}^{r} f_s(\beta_s \boldsymbol{x}_{\tau_o}) + \xi_t$$

assumes ξ_t are zero mean and independent; r is the maximum number projections or univariate regression functions; β_0 is the fixed intercept of this regression function; $\beta_s \boldsymbol{x}_{\tau_o}$ yields a (scalar) projection of the observed data as input to transformation function f_s.

This is a *fixed effects* addition to the primary univariate random effects model. We fit the projection pursuit regression onto the available `residuals' from each univariate model. We can then take as the multivariate based predicted values

$$Y_t + e_t =$$

$$\mu_{\theta_0} + \mu_{\theta_1} t(k) + \hat{\beta}_0 + \sum_{s=1}^{r} f_s\left(\hat{\beta}_s \boldsymbol{x}_{\tau_o}\right)$$

the univariate plus the PPR model to account for the error. This is a mixed effects model for prediction.

We can regard the error for the predicted values from this mixed effect model *in toto*: i.e. we do not consider the predicted values as draws from a posterior distribution (for credible intervals) but use the model based error as the variance of the predicted values. Alternately – the mean squared error of predictions from the fixed effect portion of the model can be separated from the strictly Bayesian univariate model and reported separately.

References

K Abayomi, A Gelman, M Levy “Diagnostics for Multivariate Imputation." *Journal of the Royal Statistical Society-C*. 57, Part 3, 1-19. 2008

Abayomi K., Pizarro, G. “Monitoring the United Nations Millennium Development Goals: A Simple (Bayesian) Methodology for a Cross-National Index". *Social Indicators Research*: Volume 110, Issue 2 (2013), Page 489-515

Abayomi, K., de la Pena ,V., Lall, U., and Levy, M. “Quantifying Sustainability: Methodology for and Determinants of an Environmental Sustainability Index", chapter in *Green Finance and Sustainability: Environmentally Aware Business Models and Technologies.*, ed. Z. Luo. 2010

Cao, F., Abayomi, K. and Gebraeel, N. “Probabilistic `best set' sorting algorithms for multivariate prognostic data." to appear *Algorithms and Computation* (2014).

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817–823.

Scott, A. J. and Knott, M. (1974) A Cluster Analysis Method for Grouping Means in the Analysis of Variance, *Biometrics* **30(3)**, 507–512