Research Program on Forecasting

Forecasting an Aggregate in the Presence of Structural Breaks in the Disaggregates

William Larson
Federal Housing Finance Agency
william.larson@fhfa.gov

RPF Working Paper No. 2015-002
http://www.gwu.edu/~forcpgm/2015-002.pdf

July 18, 2015

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

Forecasting an Aggregate in the Presence of Structural Breaks in the Disaggregates

William Larson*
Federal Housing Finance Agency

July 18, 2015

Abstract

There is a debate in the literature on the best method to forecast an aggregate: (1) forecast the aggregate directly, (2) forecast the disaggregates and then aggregate, or (3) forecast the aggregate using disaggregate information. This paper contributes to this debate by suggesting that in the presence of moderate-sized structural breaks in the disaggregates, approach (2) is preferred because of the low power to detect mean shifts in the disaggregates using models of aggregates. In support of this approach are two exercises. First, a simple Monte Carlo study demonstrates theoretical forecasting improvements. Second, empirical evidence is given using pseudo-*ex ante* forecasts of aggregate proven oil reserves in the United States.

JEL Codes: C52, C53, Q3
Keywords: model selection, intercept correction, forecast robustification

*I would like to thank Ryan Greenaway-McGrevy, Carlos Santos, and Tara Sinclair for their questions, comments, and encouragement throughout the research process. This work was conducted while at the Bureau of Economic Analysis. The views here are my own and not those of the Bureau of Economic Analysis or the Federal Housing Finance Agency. Please address correspondence to: William Larson, Federal Housing Finance Agency, 400 7th St SW, Washington, DC, 20024. Email: william.larson@fhfa.gov.

1 Introduction

A classic question of forecasting is whether it is best to forecast an aggregate variable directly, or to forecast the disaggregates and sum the disaggregate forecasts. The question is generally posed as one of tradeoffs: disaggregation offers the potential of greater precision, but at the risk of compounding various sources of specification errors and inefficiencies (Lutkepohl, 2006). Castle and Hendry (2010) and Hendry and Hubrich (2011) propose a third, hybrid method that attempts to capture the strengths of both of these methods without the weaknesses: forecast the aggregate directly, but include relevant disaggregate variables in the model specification. They argue that this approach is able to capture the relevant heterogeneity in the disaggregates while maintaining the efficiency of an aggregate model.

The analytical evidence in favor of Hendry and Hubrich's (2011) method is a forecast error taxonomy that compares the performance of a forecast created using an aggregate model to the performance of a forecast created by summing weighted forecasts of the disaggregates. However, this taxonomy makes one important assumption that, when relaxed, can affect their conclusions. They assume that misspecification of the long-run mean is "unlikely in both [aggregate and disaggregate] taxonomies," and consequently, there is no preference given to either approach along this dimension. However, in the periods *directly following* a mean shift, the detectability and specification of the mean shift are of fundamental concern to a forecaster. Critically, as I demonstrate in this paper, there are differences in the ability of aggregate versus disaggregated methods to detect recent mean shifts, with associated implications on model selection and forecast accuracy.

I begin by deriving a condensed version of the forecast error taxonomies of Clements and Hendry (2006) and Hendry and Hubrich (2011). The role of mean misspecification is highlighted for further examination, especially as it concerns the detectability of recent mean shifts. I analytically show that under a non-central t distribution, t-tests of the null of no mean shift have much greater power in models of disaggregates than in models of aggregates for a range of mean shift magnitudes, assuming IID innovations in the disaggregates. This low power of detection in aggregate approaches affects attempts at intercept correction (see Clements and Hendry, 1996 and Santos, Hendry, and Johansen, 2008 for discussion about intercept correction). Failure to immediately intercept correct a failed forecast has severe consequences, with resulting increases in forecast bias and RMSE in the following period. Accordingly, forecasting disaggregates and then aggregating can produce better forecasts than an aggregate model when there are moderate-sized mean shifts in disaggregates and innovations are IID.

I also examine the possibility of applying different significance thresholds α to aggregate models in hopes of increasing the power of t-tests to detect mean shifts. While higher levels of α do indeed increase rejection frequency, it is not without consequence. Inappropriate intercept correction introduces mean misspecification into the forecast. I analytically decompose the expected bias of forecasts constructed using an intercept correction strategy and find that the bias has three components: failure to properly intercept correct (bias is negative in α), contamination of the intercept correction by the current-period innovation (bias is positive in α), and false detection of a mean shift (bias is positive in α). This decomposition allows for levels of α to be simulated that are optimal in the sense that they minimize expected bias given various criteria. In the example provided, in a model of a disaggregate, the median bias-minimizing level of α is 0.0005 and the 90th percentile bias-minimizing level of α is 0.005. Increasing α above 0.005 increases expected bias due to the possibility of Type I error. On the other hand, in the aggregate-forecast example, the median bias-minimizing level of α is 0.13—much higher than in the disaggregate case—suggesting that increased mean shift detection sensitivity may be desirable in models of aggregates. The bias-minimizing 90th percentile strategy is to maintain the stricter α threshold of 0.003. In all cases, the disaggregate modeling approach weakly out-performs the aggregate modeling approach in terms of median bias.

Next, I empirically illustrate these concepts by considering the problem of forecasting proven U.S. oil reserves. Oil reserves come close to an ideal illustration because innovations are mostly IID, the aggregation weights are known and constant over time, and the disaggregates are subjected to orthogonal and unanticipated mean shifts. Mean shifts occur due to a variety of reasons, including drilling innovations, regulation, and discovery. I initially consider three methods of forecasting aggregate reserves: forecasting disaggregates and then aggregating, forecasting the aggregate directly using only aggregate information, and forecasting the aggregate using both aggregate and disaggregate information using an *Autometrics* algorithm (Doornik, 2009) with Impulse Indicator Saturation (IIS; Santos, Hendry, and Johansen, 2008). I then also consider two methods of forecast combination: a simple average of forecasts and a switching forecast that is motivated by analytical predictions from the prior sections.

The disaggregate forecasting method easily detects mean shifts with very little Type 1 error, making it the preferred method during turbulent periods, but somewhat less efficient than the aggregate methods. The aggregate methods are also able to detect mean shifts, but with greater risk of Type 1 error. One alternative to either of these individual approaches is a

switching method. First, a pre-test for mean shifts on each of the disaggregates is conducted. Then, if a moderate number of breaks are detected, the disaggregate approach is used to forecast the period. Alternatively, if either a small or large number of breaks are detected, one of the more parsimonious aggregate approaches is used. This switching forecast is based on the analytical result that when mean shifts in disaggregates are either very small or very large, then power differentials of tests of mean shifts between aggregate and disaggregate approaches is small. On the other hand, when breaks are moderate in size, the power differential is large. This switching forecast takes advantage of the best attributes of each of the forecasting approaches: when times are uncertain, the more resilient, disaggregate-based approach is used; when times are more certain, then the aggregate approach is more parsimonious and therefore more efficient.

2 Assumptions and Forecast Error Taxonomy

This section introduces a simple analytical framework in order to examine some of the important sources of forecast error. I begin with a forecast error taxonomy that decomposes sources of forecast error into various components, following Clements and Hendry (2006) and Hendry and Hubrich (2011). This taxonomy is derived for both disaggregate and aggregate approaches of forecasting an aggregate series. Much of this taxonomy is a condensed version of Hendry and Hubrich (2011), but with one key departure, the lack of an autoregressive term in the hypothetical DGP. This simplifies the analysis and allows for greater focus on the effects of unanticipated mean shifts in the disaggregates.

Of particular interest is the effect of misspecification of the mean across forecasting approaches. Hendry and Hubrich (2011) suggest mean misspecification to be unlikely asymptotically and with no large differences between methods. While the mean is constant, this is likely the case. In a period where the mean shifts, the cause of forecast error in that period is indeed not mean misspecification—breaks are impossible to forecast and therefore are unmodelable. Rather, the source of error is the mean shift itself, which exists as a separate error cause in the taxonomies.

Where things become more complicated are the periods *directly following* a mean shift. The shift exists in-sample, and therefore must be incorporated into any estimate of the mean. This poses a serious problem: what methods are able to detect and incorporate mean shifts quickly and accurately? The accuracy of the forecast hinges on the ability of the forecaster to detect the mean shift and appropriately specify the forecasting model. This problem has important consequences regarding forecast accuracy across different forecasting approaches.

Assumptions

Suppose there are n disaggregates in the vector Y_t, with each element denoted Y_{it}. The weighted sum of Y_t is defined as the aggregate Y^a. The weights are known and do not change over time.

$$Y_t^a \equiv \omega' Y_t \tag{1}$$

Each disaggregate is I(0) with a long-run mean and a structural innovation each period, with the vector of each denoted by δ and ϵ_t, respectively. The structural innovations are drawn from the distribution Ω.

$$Y_t = \delta + \epsilon_t \quad \text{for } t = 1, ..., T \tag{2}$$

$$\text{where } \epsilon_t \sim ID(0, \Omega)$$

There is a break in the vector of means at T that is unknown to the forecaster, giving

$$Y_{T+h} = \delta^* + \epsilon_{T+h} \quad \text{for } h = 1, ..., H \tag{3}$$

The new DGP is identical in all respects except for this mean shift. The question now turns to examining sources of forecast error, and how they are similar or dissimilar depending on the forecasting approach used.

Forecast Error Taxonomy

In this first exercise, the forecaster models Y for each disaggregate. This model is used to construct forecasts $\hat{Y}_{iT+1|T}$, which give the aggregate forecast $\hat{Y}^a_{T+1|T} \equiv \omega' \hat{Y}_{T+1|T}$. The resulting forecast error is

$$\omega' \hat{\epsilon}_{T+1|T} = \omega' \delta^* - \omega' \hat{\delta} + \omega' \epsilon_{T+1} \tag{4}$$

The disaggregate mean vector $\hat{\delta}$ is nx1 and is asymptotically distributed around δ_{dis} according to $\sqrt{T}\hat{\delta} \sim (\delta_{dis}, \Omega/T)$ giving estimation error equal to $\delta_{dis} - \hat{\delta}$. The model used to estimate $\hat{\delta}$ may also be misspecified, introducing bias equal to $\delta - \delta_{dis}$. These concepts of model misspecification and estimation error can be incorporated into Equation 4 by adding and subtracting both $\omega'\delta$ and $\omega'\delta_{dis}$, giving rise to a forecast error taxonomy similar to those in Clements and Hendry (2006) and Hendry and Hubrich (2011).

$$
\begin{aligned}
\omega'\hat{\epsilon}_{T+1|T} = & & (5)\\
&= \omega'(\delta^* - \delta) && (a) \text{ unanticipated mean change}\\
&+ \omega'(\delta - \delta_{dis}) && (b) \text{ mean misspecification}\\
&+ \omega'(\delta_{dis} - \hat{\delta}) && (c) \text{ estimation error}\\
&+ \omega'\epsilon_{T+1} && (d) \text{ stochastic innovation}
\end{aligned}
$$

Forecasting Aggregates Directly

Alternatively, a forecaster can model Y^a directly using a single, aggregate model. This model is used to directly construct the forecast $\tilde{Y}^a_{T+1|T}$. In this case, the resulting forecast error is

$$\tilde{\epsilon}_{T+1|T} = \omega'\delta^* - \tilde{\delta} + \omega'\epsilon_{T+1} \tag{6}$$

The estimate of the disaggregate mean $\tilde{\delta}$ is now 1x1, and distributed around δ_{agg} according to $\tilde{\delta} \sim (\delta_{agg}, \omega'\Omega\omega/T)$. This gives the forecast error taxonomy

$$
\begin{aligned}
\tilde{\epsilon}_{T+1|T} = & & (7)\\
&= \omega'(\delta^* - \delta) && (A) \text{ unanticipated mean change}\\
&+ \omega'\delta - \delta_{agg} && (B) \text{ mean misspecification}\\
&+ \delta_{agg} - \tilde{\delta} && (C) \text{ estimation error}\\
&+ \omega'\epsilon_{T+1} && (D) \text{ stochastic innovation}
\end{aligned}
$$

Comparing Sources of Forecast Error

The following comparison of the two previously established forecast error taxonomies provides the basis for the remainder of the paper. First, it is clear that (a) and (A), as well as (d) and (D), are identical. Thus, no differences in forecast performance can be attributed to either the mean shift or the stochastic innovation. This is not surprising, given that both are impossible to predict *ex ante*. Differences therefore must arise due to differences in specification and estimation error.

Differences in Estimation Error

Estimation error differences hinge on the structure of the joint distribution Ω from which innovations are drawn, the aggregation weights ω and the sample size T. As Lutkepohl (2006) demonstrates, asymptotically, the disaggregate approach is superior due to the elimination of estimation error in the disaggregate parameters. However, in small samples, the compounding of estimation error in the disaggregates can cause problems such that a more parsimonious approach is superior. This is the "bias-variance" tradeoff often mentioned in the forecast aggregation literature (see Friedman, 1997 for one of the many discussions on the subject). In general, as Hendry and Hubrich (2011) conclude, "it is not possible to make general statements about whether differences in forecast accuracy are mainly due to the bias or the variance of the forecast."

Differences in Mean Misspecification

Mean misspecification is primarily a model selection issue. While Hendry and Hubrich (2011) note, "long run mean misspecification...is unlikely when the in-sample DGP is constant and the model is well-specified," this only statement is only valid in the periods leading up to and including the break period T. Up until time $T+1$, the mean can usually be modeled with a constant term that does not change over time. In period $T+1$, the break is an unanticipated mean shift under the classification (a) or (A) in the above taxonomies. In periods following $T+1$, the mean shift has been observed and is now in the data, requiring it to be properly modeled in order to avoid forecast failure.

Correctly specifying the model when there is a recent mean shift is no small task, as the DGP has undergone a dramatic change. The "long run" is instead very much the "short run," and issues of small-sample statistical testing such as test power become relevant. There are various strategies of detecting and modeling mean shifts, which will be investigated beginning with the next section.

3 The Power of Test Statistics

The problem turns to the detection of mean shifts using small-sample statistical tests. This involves establishing the power of tests evaluating the null of no mean shift when a mean shift has in fact occurred. Because the null is known to be false, the test statistic follows a non-central t distribution. In this section, the power of t-tests under a non-central t-distribution are considered for both the disaggregate and aggregate modeling approaches.

In order to further simplify the analysis, some restrictions are placed on Ω and ω. First, Ω is now assumed diagonal, with the same variance σ^2 in each disaggregate, and with normally distributed structural innovations drawn such that $\epsilon_{it} \sim IN(0, \sigma^2)$. Second, $\omega_i = 1$ for all i, which is the case when the aggregate is simply the sum of the disaggregates. Combined, these assumptions make differences in estimation error (taxonomy categories (c) and (C)) equal to zero, isolating the effects of mean misspecification, (b) and (B) in Equations 5 and 7.[1] Under these conditions, as this section shows, tests of mean shifts have lower power in the aggregate approach versus the disaggregate approach.

Suppose δ is equal to zero until time T, at which the first disaggregate's mean shifts to d.[2]

$$\delta_{it} = \begin{cases} 0 & \text{for } t = 1, ..., T \\ d & \text{for } t = T+1, ..., T+H \text{ and } i = 1 \end{cases}$$

The forecaster tests for breaks each period in order perform intercept correction in future periods. Intercept correction, which is covered in greater detail in the next section, is a method of forecast robustification advocated by Clements and Hendry (1996) which involves correcting a "missed" forecast by an amount related to the forecast error if the error (or average of past errors) exceeds a certain threshold level of statistical significance. Here, this requires selecting a threshold level of significance α and testing the following parameter

$$\hat{\rho}_{it} = \delta_{it} + \epsilon_{it} \tag{8}$$

where $E[\hat{\rho}_{it}] = \delta_{it}$ and $V[\hat{\rho}] = \sigma^2$. When the null is known to be false, the t-statistic follows a non-centered t distribution with a non-centrality parameter ψ. In the case of $\delta_{1T} = d$,[3]

$$E\left[t_\rho(\psi)\right] \simeq \frac{\delta}{\sigma} \equiv \psi \tag{9}$$

[1] The aggregate stochastic innovation is the same in both cases and equal to $n\sigma^2$. In the disaggregate model, asymptotically, $\sqrt{T}\hat{\mu}_i \sim IN(0, \sigma^2/T)$, so summing over n disaggregates, $\sum \hat{\mu} \sim IN(0, n\sigma^2/T)$. Equivalently, in the aggregate model, $\sqrt{T}\hat{\mu} \sim IN(0, n\sigma^2/T)$. The resulting difference in forecast error variance is thus the difference in the squared biases alone.

[2] The notation and sequence of derivations is based on Santos (2008).

[3] $t_\rho(\psi) = \frac{\hat{\rho}}{\hat{\sigma}} = \frac{\frac{\hat{\rho}-\delta}{\sigma} + \frac{\delta}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{(n-1)\hat{\sigma}^2}{\sigma^2}}}$ and $E\left[\frac{\frac{\hat{\rho}-\delta}{\sigma} + \frac{\delta}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{(n-1)\hat{\sigma}^2}{\sigma^2}}}\right] \simeq \frac{\delta}{\sigma}$.

The power of the test p is the probability P of rejecting the null

$$p_\psi = P\left[t_\rho(\psi) > c_\alpha | \psi \neq 0\right] \tag{10}$$

Now, consider the aggregate which is the sum of the disaggregates.

$$Y_t^a = \delta_t^a + \epsilon_t^a \tag{11}$$

where $\delta_t^a \equiv \sum \delta_{it}$ and $\epsilon_t^a \equiv \sum \epsilon_{it}$. Similarly, other superscripts a denote aggregates. Testing for a break involves testing the significance of the parameter

$$\hat{\rho}_t^a = \delta_t^a + \epsilon_t^a \tag{12}$$

where $E[\hat{\rho}_t^a] = \delta_t^a$ and $V[\hat{\rho}_t^a] = n\sigma^2$. Note that Equations 8 and 12 are nearly identical, except the error variance is n times larger in the aggregate. The t-statistic is therefore $1/\sqrt{n}$ as large as in the disaggregate case, with associated lower power.

Assuming $\sigma^2 = 1$ and $n = 10$ gives the following power curves during the period of the break, estimated numerically using STATA's NCT function (Steichen, 2000). First, looking at Figure 1a, it is observed that for small mean shifts in Y, it is difficult to reject the null of no mean shift, so the test has low power at all significance levels. The higher the mean shift or the lower the significance level, the higher the power of the test. For instance, 80% power is achieved when $d = 2$ and $\alpha = 0.1$, whereas the $\alpha = .001$ test requires a $d = 4$ for the same power.[4] When comparing the aggregate to the disaggregate in Figures 1b and 1c, it is clear that the power of the test is much lower in the aggregate because of the higher error variance at moderate ranges of mean shifts. For either small or large mean shifts, there is little difference between the power of tests using the two approaches.

Figure 1: Power of t-tests under a non-central t distribution

4 Intercept Correction using Significance-Based Decision Rules

Intercept correction (IC) is a method of forecast robustification advocated by Clements and Hendry (1996). In this method, suppose a forecast error $\hat{e}_t$. Then, the value $\hat{e}_t$ is added

[4]Obviously, high power due to low significance thresholds is not without costs. Such a decision rule invites false rejections of a true null, with resulting additions to forecast error. Consideration of the "optimal" value of α will be considered in later sections.

to each forecast from $t+1$ onward, and this approach is repeated for every t. Clements and Hendry's (1996) "always IC" method quickly gets a forecast that has failed due to a deterministic mean shift back onto the correct path, but at the risk of increasing the noise in the forecast in cases where the forecast error is not due to a mean shift.

Alternatively, it is possible to create a decision rule governing the IC process. This rule is generally formed based on a chosen significance level α of a residual or dummy variable in the final forecasting period before a new forecast is generated.[5] If the dummy/residual is statistically significantly different than zero, then IC is adopted. This framework is often implemented using impulse indicator saturation (IIS) in an Autometrics model selection algorithm (see Santos, 2008; Santos, Hendry, and Johansen, 2008; Doornik, 2009; and Castle, Hendry, and Doornik, 2012).

Typically, these significance levels are somewhat arbitrarily set by convention based on desired Type I error. However, when it comes to the practice of IC adoption, the chosen α has real consequences on the accuracy of forecasts (see Castle, Fawcett, and Hendry, 2011). In the prior section, I establish that detection of mean shifts is more difficult in aggregate models. Therefore, it is plausible that lowering the significance threshold by choosing a higher α can increase forecasting performance in aggregate models. It is also likely that lowering the significance threshold increases the chances of false rejection of the null of no mean shift (Type I error), as the initial Clements and Hendry (1996) "always IC" approach does.

These issues of mean shift detection and intercept correction in models of aggregates versus disaggregates have not been adequately considered in the literature. As the remainder of this section shows, there are various sources of Type I and Type II error that must be considered, and each has different effects on expected bias.

Continuing from the prior example, the following is the expected forecast for the aggregate Y_i in period $T+1$ for the disaggregate, where $p_{\alpha,\psi,\hat{\delta}}$ is the probability of retaining a dummy variable with magnitude $\hat{\delta}$ under a non-central t distribution with non-centrality parameter ψ (if $\psi = 0$, the distribution follows the Student's t distribution) and a significance level α.

[5] Alternatively, one can use a weighted average of forecasts, such as performed by Castle et al. (2009). Another option is to follow Hendry and Santos (2005), who argue that there are gains from "last sample observation indicators" that are not extrapolated to future periods. These are less risky in the sense they merely omit the final observation from the model, which often is measured with greater error than other observations in the sample, as opposed to taking the error and projecting it forward.

There is a break in $i = 1$ of size $d = \psi\sigma$.

$$E[\hat{Y}^a_{T+1|T}] = p_{\alpha,\psi,\hat{\delta}_1}\hat{\delta}_1 + \sum_{i=2}^{I} p_{\alpha,0,\hat{\delta}_i}\hat{\delta}_i \tag{13}$$

Substituting $\hat{\delta}_1 = d + \epsilon_{1T}$, $\hat{\delta}_{i\neq 1} = \epsilon_{iT}$, and $\psi = d/\sigma$ gives

$$E[\hat{Y}^a_{T+1|T}] = p_{\alpha,d/\sigma,d+\epsilon_{1T}}(d + \epsilon_{1T}) + \sum_{i=2}^{I} p_{\alpha,0,\epsilon_{iT}}\epsilon_{iT} \tag{14}$$

The expected value of $\hat{Y}^a_{T+1}$ when the mean shift is known is simply $E[\hat{Y}^a_{T+1}] = d$. The bias of this forecast is the mean misspecification according to Equation 5 (b), defined as $\hat{b}_{T+1|T} \equiv Y^a_{T+1} - \hat{Y}^a_{T+1|T}$, and the expected bias for the disaggregate approach is therefore

$$E[\hat{b}_{T+1|T}] = \underbrace{(1 - p_{\alpha,d/\sigma,d+\epsilon_{1T}})d}_{\text{Failure to IC}} - \underbrace{p_{\alpha,d/\sigma,d+\epsilon_{1T}}\epsilon_{1T}}_{\text{IC Contamination}} - \underbrace{\sum_{i=2}^{I} p_{\alpha,0,\epsilon_{iT}}\epsilon_{iT}}_{\text{False IC}} \tag{15}$$

Here, we can see the tradeoff in choosing a significance level α. When α is high, the probability of rejecting the null of no mean shift is also high. Bias falls to zero in the first term as $1 - p \to 0$. On the other hand, the second and third terms increase with α. Error introduced by the second term may be desirable in the case of a mean shift when the error and the shift are of the same sign but the mean shift is hard to detect. While introducing bias by contaminating the IC, in this case, there is less bias than not detecting the break. The third term is the cost of performing intercept correction when in fact no mean shift has occurred.

Similarly, for the aggregate, defined as $\tilde{b}_{T+1|T} \equiv Y^a_{T+1} - \tilde{Y}^a_{T+1|T}$,

$$E[\tilde{b}_{T+1|T}] = \underbrace{(1 - p_{\alpha,d/(\sqrt{I}\sigma),d+\epsilon_{1T}})d}_{\text{Failure to IC}} - \underbrace{p_{\alpha,d/(\sqrt{I}\sigma),d+\epsilon_{1T}}\epsilon_{1T}}_{\text{IC Contamination}} \tag{16}$$

Note that the noncentrality parameter falls from d/σ to $d/(\sqrt{n}\sigma)$ in the aggregate case, reducing the power of tests of mean shifts.

The functional forms of p are quite complicated as shown in the prior section, so numerical representation of Equations 15 and 16 are shown below. The numerical exercise begins with one of ten disaggregates facing a mean shift of $d = 4$ standard deviations in the 100th period. The ϵ_is are then independently drawn and hypothesis testing is performed for each

$\alpha = \{0.000, 0.005, ..., 0.150\}$. Intercept correction is then performed when the null of no break is rejected, and expected bias is computed following Equation 15. Simulated bias is presented below for 10,000 replications, first for the disaggregate and then the aggregate approach. The black line is the median and shaded areas are successive 10% bands.

Figure 2: Expected Bias of Forecasts at Different Levels of α

In Figure 2a, at $\alpha = 0$, the expected bias is $d = 4$ because there is no intercept correction at any threshold. However, this bias can be lowered even with very low levels of α, with the minimum median bias estimate is achieved at $\alpha \approx 0.0005$, while the minimum 90% error occurs at $\alpha \approx 0.0035$. Past this value, the benefits of increased detection of the break are more than offset by the danger of intercept correcting when in fact there was no break.[6] On the other hand, the median bias from the aggregate approach (shown in Figure 2b) is monotonically decreasing in α, while the 90% confidence interval of the bias slowly increases due to IC contamination.

Figure 3: Difference in Median Bias, Aggregate Approach minus Disaggregate Approach

5 An Illustration: Forecasting U.S. Proven Oil Reserves

In this section, I take the role of a forecaster attempting to produce pseudo *ex ante* forecasts of U.S. aggregate proven (or "proved") oil reserves in the 2008-2012 period. This variable is ideal to illustrate the concepts in the prior section because its innovations are mostly IID in the disaggregates and it is subject to unanticipated mean shifts due the recent discovery and widespread adoption of hydraulic fracturing and horizontal drilling techniques.

There are 47 different distinct areas with oil reserves, which make up the entirety of the total U.S. reserves. These mutually exclusive regions can consist of states, sub-regions within large states (such as Texas), or offshore areas (areas in the Gulf of Mexico or off of the coast of California). Figure 4 shows U.S. proven oil reserves and Cushing, Oklahoma spot crude oil prices from 1986-2012. Reserves had been previously trending downwards from 1986-2008, then abruptly reversed course from 2009-2012. The question which motivates

[6] It should be emphasized that these simulations consider the case where the innovations are IID. Non-diagonal covariance matrices will have varying power differences. For example, if correlations between innovations are negative, then innovations in disaggregates will cancel in the aggregate, increasing the power to detect mean shifts using aggregate approaches. Alternatively, if the correlations between innovations are positive, then the power of tests using aggregate approaches decreases.

this forecasting exercise is when and how could this mean shift in the growth rate of proven oil reserves be detected?

Figure 4: Proven Oil Reserves and Prices, 1986-2012

Following Farzin (2001), proven oil reserves are modeled as a function of extraction rates μ, mean shifts d, oil prices, and IID innovations. Prices affect proven oil reserves for two reasons: first, prices incentivize oil search and discovery which is costly; and second, prices affect reserves by definition because reserves are the quantity which, "with reasonable certainty, are recoverable under existing economic and operating conditions," according to the U.S. Energy Information Administration. The mean shifts in the growth rate are location-specific, and can be positive (i.e. horizontal drilling technology) or negative (i.e. Arctic National Wildlife Refuge exploration restriction). The stochastic specification of this model is

$$\Delta res_{it} = \mu_i + d_{it} + \beta \Delta price_t + \epsilon_{it} \tag{17}$$

where d consists of an unknown, small number of breaks for each cross-sectional unit. A Dickey-Fuller test cannot reject the null of a unit root in the log-level of oil prices (both with and without a trend) at any significance level, and rejects the null at the 1% level in the first-difference, suggesting oil prices are an I(1) process. Lag selection by AIC and SBIC each gives a lag length of zero. Therefore, oil prices are modeled as a random-walk-with-drift process.

$$\Delta price_t = \lambda + \varepsilon_t \tag{18}$$

The break timings are unknown to the forecaster but the forecaster knows breaks sometimes occur, so a detection strategy is devised. Intercept correction is performed for all future periods when one of the following conditions is met: 1) the current period model residual exceeds a certain threshold level of significance based on a chosen significance level α; 2) if the most recent two residuals are statistically indifferent from each other and different than zero; and finally, 3) if the most recent three residuals are statistically indifferent from each other and different than zero. Case (1) is obvious as classic intercept correction, but methods (2) and (3) allow for intercept correction if a smaller break exists for two or three periods, respectively. These two additional cases ensure that breaks are incorporated into future forecasts eventually, though sometimes it takes multiple periods to be sure of a break.

This forecasting strategy is then implemented using three modeling approaches. In all

cases, 1-step forecasts of the growth rate of the log-difference are estimated conditional on the estimated price drift $\hat{\lambda}$.[7] Then, the forecasted level is created by multiplying the log-difference by the prior period's level.

In the first of three approaches, forecasts are constructed using the aggregate model with no disaggregate information. In the second, the disaggregates are modeled and forecasts are constructed for the disaggregates, and the aggregate forecast is created by summing the disaggregate forecasts. Finally, an Autometrics routine is implemented following Doornik (2009) using an aggregate model, but with a full information set of aggregate and lagged disaggregate elements. In this third method, impulse indicator saturation (IIS) is performed as well (see Castle, Doornik, and Hendry, 2012). The use of lagged disaggregates gives equivalent 1-step forecasts to a VAR, which is what Hendry and Hubrich (2011) use to incorporate disaggregate information. Forecasts are then constructed based on the model chosen from this algorithm.

Aggregate-Only Forecast

Figure 5 shows recursive 1-step forecasts using the aggregate-only approach. The blue line is the actual, the red shows aggregate forecasts with no intercept correction, and the green line shows forecasts with intercept correction if p-value of the t-test of the residual (or residuals) is less than $\alpha/2$. In Panel 5a, the decision to intercept correct is made in 2010 based on large, similar residuals for 2009 and 2010. Accordingly, the forecasts for 2011 and 2012 are relatively accurate. On the other hand, in Panel 5b, 2008 is determined to be in need of intercept correction, with resulting forecast failure in 2009. This case is representative of the danger of Type I error in intercept correction. On the other hand, the more liberal threshold gives quick adjustment to the miss in 2009 and provides forecasts that are back on track for 2010 on. Panels 5c and 5d are similar to but noisier than Panel 5b, suggesting that a more stringent criteria in the $\alpha = [0.0001, 0.001]$ range is preferable.

Figure 5: Forecasts using Aggregate-Only Models

[7] 1-step forecasts are considered because at longer time horizons, the test power differential between the aggregate and disaggregate approaches falls with the length of the forecast horizon. I leave the derivation of forecast power differentials as a function of forecast horizon length to further research.

Aggregating Forecasts of Disaggregates

Figure 6 shows recursive 1-step forecasts constructed using the disaggregated approach. As before, the blue line is the actual, the red shows aggregate forecasts with no intercept correction (but constructed using the disaggregate approach), and the green line shows forecasts with disaggregated intercept correction. In Panel 6a, a mean shift is detected in 2008 in 4 of the 47 disaggregates. These corrections, along with 7 breaks in 2009 result in a small miss in 2009. However, because of the new breaks detected in 2009, the forecasts for 2010-2012 are back on track.

Successive weakening of α serves to make the disaggregates more sensitive to intercept correction, but unlike the aggregate case, this reduces forecast accuracy. The forecasts for 2009 become less accurate due to the greater number of false breaks detected in 2008 using the lower acceptance thresholds. Perhaps most alarmingly, the pre-2008 forecasts become noisier as α increases. Combined, these results suggest $\alpha = 0.0001$ to be the preferred threshold for intercept correction.

Figure 6: Forecasts using Disaggregate Models

Aggregate Model using Disaggregate Information

Figure 7 shows recursive 1-step forecasts constructed using the Autometrics model selection algorithm (see Doornik, 2009) on an information set consisting of each disaggregate and a full set of time dummy variables with variable retention significance threshold α. This incorporates Hendry and Hubrich's (2011) notion of capturing disaggregate information, and improves on it by employing a model selection algorithm using a general-to-specific (GETS) approach. The method here encompasses the aggregate-only with intercept correction strategy in Section 5 because similar decision rules governing the selection of dummy variables in the Autometrics algorithm are used in the decision to intercept correct large residuals.

Higher levels of α tend to perform better than low levels, but all appear to produce somewhat noisy forecasts. The algorithm retains some unusual variables with questionable signs. For example, West Virginia, with only about 0.1% of U.S. reserves, has a significant, *negative* effect on U.S. reserves in nearly every year's model when $\alpha = 0.1$ or $\alpha = 0.01$. Thresholds stricter than $\alpha = 0.001$ do not retain a constant term in most models. Despite these odd correlations (or lack thereof) in the final models, the forecasts tend to be fairly

accurate on average.[8]

Figure 7: Forecasts using Aggregate Models with Disaggregate Information

Comparison of Forecasts

Figure 8 shows the best forecasts of each of the prior methods. Pre-2009, the disaggregate forecasts perform slightly worse than the aggregate forecast. Despite the theoretical superiority of the GETS algorithm combined with a large information set, the forecasts produced by the resulting model appear to be noisier. In 2009, we see how much the Type I error from the 2008 mean shift detection reduces forecast accuracy. In the aggregate model case, the large residual in 2008 is misinterpreted as a break and causes forecast failure in 2009, whereas this does not occur in the disaggregate approach. However, for 2010, the 2009 mean shift is captured almost perfectly by the 2009 intercept in the aggregate model, whereas it is only partially accommodated in the disaggregate estimates, and not at all using the aggregate model with disaggregate information. In general, the median error appears is lower in the aggregate model, but with a greater possibility of forecast failure. Disaggregation appears to act as insurance, not just against large breaks, but against Type I error in the detection of breaks as well.

Figure 8: Best Forecasts using Alternative Modeling Approaches

Based on the results from the bias exploration in Section 4, a switching decision rule is devised in hopes of incorporating the strengths of the approaches. Recall that for both high and low magnitude mean shifts, there is little power differential in the detection of mean shifts in aggregate versus disaggregate models. Accordingly, there is little to gain by disaggregating when mean shifts are either non-existent or large. Figure 9 shows the fraction of areas (both nominally and weighted by fraction of total reserves) where a mean shift is detected at the $\alpha = 0.0001$ significance level.

[8]These noisy forecasts may be due to omitted variable bias in the estimates of the other parameters. This is important because in the hybrid method that they propose—forecasting aggregates using additional disaggregate information—estimates are especially vulnerable to this omitted variable bias. If a disaggregate right-hand-side variable model-encompasses other disaggregates prior to the break, then the information in the variable is proxying for information in other disaggregates. When this correlation becomes nonconstant by means of an unmodeled mean-shift, the resulting error is projected onto the other disaggregates for which it was a proxy.

Figure 9: Fraction of Disaggregates with Identified Mean Shifts in Particular Year

A moderate number of (weighted) breaks appears in 2003, 2004, and 2008, with a large number of breaks in 2009. All other periods have very low (or zero) numbers of breaks. Based on this, a "switching forecast" is generated using forecasts from the disaggregate approach with intercept correction in 2003, 2004, and 2008, and using the aggregate approach with intercept correction in all other periods. Given that this is a weighted average of forecast with weights equal to one or zero, I also present a forecast with equal weights (50% to each). Both of these forecasts perform quite well, as shown in Figure 10.

Figure 10: Forecast Combinations: Switching and Average Forecasts

The following table of RMSEs shows the GETS algorithm along with impulse indicator saturation, despite its sometimes unintuitive model estimates, forecasts better than any other aggregate model. Intercept correction is also demonstrated to be of critical importance to getting a forecasting model back on track after a mean shift, as demonstrated by the superiority of the IC forecasts relative to the no IC forecast. The disaggregate approach with intercept correction shows that in the case of proven oil reserves in the United States, the power of test statistics to detect mean shifts is of fundamental importance. The switching model produces slightly better point forecasts than any single model.

Table 1: RMSE of Different Forecasts

Forecast	RMSE
Aggregate Model, no IC	2.56%
Aggregate Models, IC $\alpha = 0.001$	2.16%
Aggregate with Disagg. Information, Autometrics with $\alpha = 0.001$	1.79%
Disaggregate Models, IC $\alpha = 0.0001$	1.65%
Average Forecast	1.85%
Switching Forecast	1.56%

6 Conclusions

In this paper, I examine some of the issues fundamental to the choice of methods when the goal is to forecast an aggregate. Building on prior work by Clements and Hendry (2006), Hendry and Hubrich (2011), and many others who address this topic, I consider the possibility that unanticipated mean shifts occur in disaggregates, and the conditions under which

they are quickly detectable in aggregate versus disaggregate modeling frameworks. This relaxes a fundamental assumption in Hendry and Hubrich (2011), who assume that mean misspecification is small in both aggregates and disaggregates, and can be largely ignored. While this is true in the long run, in the periods directly following a mean shift, the specification of the mean is actually a very important problem, as the failure to incorporate a mean shift into a forecasting model is a major source of forecast failure.

In the analytical section, I begin by first showing that there exists a substantial differential in the power of non-central t-tests of the null of no mean shift when modeling disaggregates individually versus an aggregate. This is consequential when these tests are part of intercept correction decision rules (Clements and Hendry, 1996; and Castle, Doornik, and Hendry, 2012). Because they are consequential and there is a differential in detection power, I explore the possibility of different significance level thresholds for mean shift detection. Using a Monte Carlo simulation, I show that the optimal significance threshold for intercept correction, α, is less in a model of a disaggregate than in a model of an aggregate, and is very low ($\alpha = [0.0001, 0.001]$) in the example considered.

This conclusion regarding α is based, in part, on a bias decomposition that is helpful to understand sources of forecast failure following a mean shift. There are three main sources of bias: first, failure to detect a mean shift results in no intercept correction (bias decreases in α). Second, a detected mean shift is impossible to distinguish from innovations in the period, leading to contamination of the intercept correction (bias increases in α). Finally, in disaggregate approaches, there is the possibility of Type I error, or the false detection of a mean shift (bias increases in α).

In the empirical section, I present a real-world illustration of each of the types of error in the bias decomposition in the prior section. U.S. proven oil reserves exhibit characteristics that are ideal for this sort of analysis because new discoveries, technologies, and regulations are, for the most part, independent. Over the 2000-2012 time period, there are large residuals and structural mean shifts in both disaggregates and aggregates. Disaggregate modeling facilitates detection of these mean shifts with less Type I and II error. On the other hand, disaggregation is costly in the sense that it increases the noisiness of the aggregate forecast. Aggregate models are much more parsimonious and produce better forecasts during periods without breaks. However, it is much more difficult to detect breaks, leading to both Type I and Type II errors in their identification.

The superior forecast is based on the following forecasting procedure. First, disaggregates are tested for breaks. Then, if no breaks or a large number of breaks are detected, an

aggregate model is used to forecast. On the other hand, if moderate number of breaks are detected, then each disaggregate is forecasted and the results are summed. This approach is consistent with the analytical prediction that mean shift detection is superior in disaggregates when the mean shifts are of a moderate size, but not too small (neither approach can detect) or large (both can easily detect). In this particular case, this strategy out-performs both wholely disaggregate and aggregate approaches, including the approach advocated by Hendry and Hubrich (2011) of incorporating disaggregate information into aggregate models. This is operationalized by modeling an aggregate using a full information set of the disaggregates along with impulse indicator saturation (Santos, Hendry, and Johansen, 2008) in Autometrics (Doornik, 2009).

References

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2012). Model selection when there are multiple breaks. *Journal of Econometrics*, 169(2):239 – 246. Recent Advances in Nonstationary Time Series: A Festschrift in honor of Peter C.B. Phillips.

Castle, J. L., Fawcett, N. W., and Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, 210(1):71–89.

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2011). Forecasting breaks and forecasting during breaks. Working Paper 535, University of Oxford Department of Economics.

Castle, J. L. and Hendry, D. F. (2010). Nowcasting from disaggregates in the face of location shifts. *Journal of Forecasting*, 29(1-2):200–214.

Clements, M. P. and Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics*, 11(5):475–494.

Clements, M. P. and Hendry, D. F. (2006). Forecasting with breaks. volume 1 of *Handbook of Economic Forecasting*, pages 605 – 657. Elsevier.

Doornik, J. A. (2009). Autometrics. In *In Honour of David F. Hendry*, pages 88–121. University Press.

Farzin, Y. (2001). The impact of oil price on additions to US proven reserves. *Resource and Energy Economics*, 23(3):271 – 292.

Friedman, J. H. (1997). On bias, variance, 0/1 loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77.

Hendry, D. F. and Hubrich, K. (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business and Economic Statistics*, 29(2):216–227.

Hendry, D. F. and Santos, C. (2005). Regression models with data-based indicator variables. *Oxford Bulletin of Economics and statistics*, 67(5):571–595.

Lutkepohl, H. (2006). Forecasting with VARMA models. Handbook of Economic Forecasting, pages 287 – 325. Elsevier.

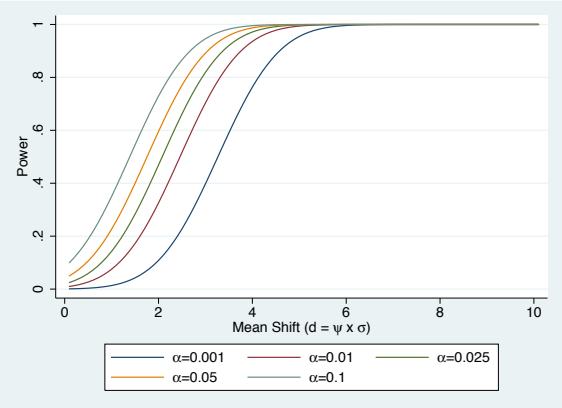
Santos, C. (2008). Impulse saturation break tests. *Economics Letters*, 98(2):136 – 143.

Santos, C., Hendry, D. F., and Johansen, S. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 23(2):317–335.

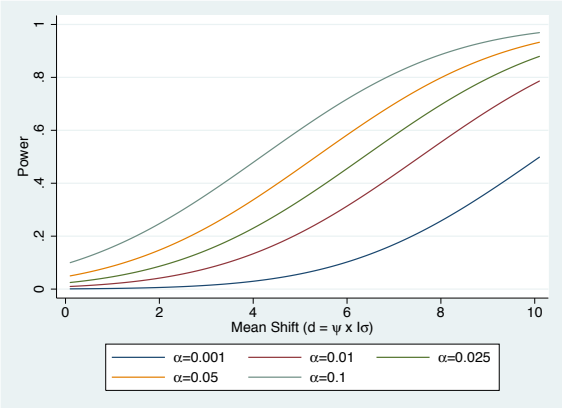
Steichen, T. (2000). NCT: Stata modules related to the noncentral t distribution. Boston College Department of Economics.

Figure 1: Power of t-tests under a non-central t distribution

(a) Disaggregate



(b) Aggregate



(c) Difference

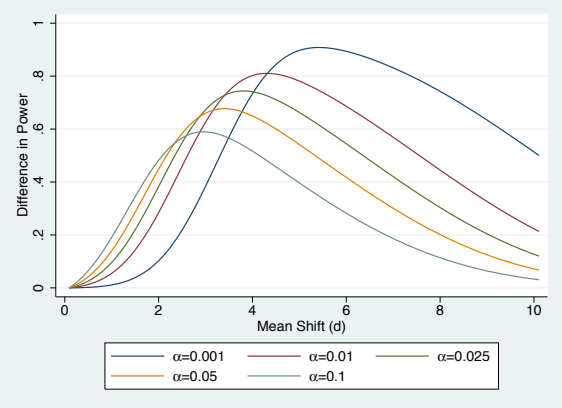
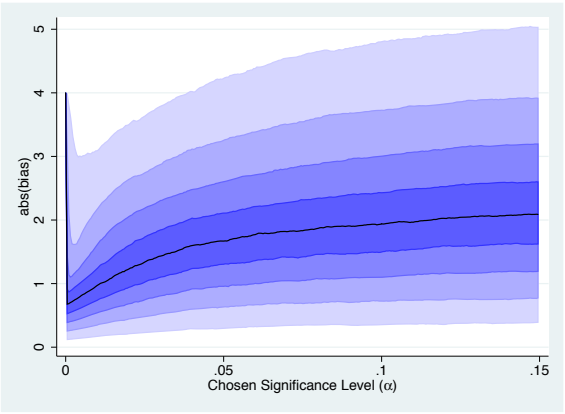
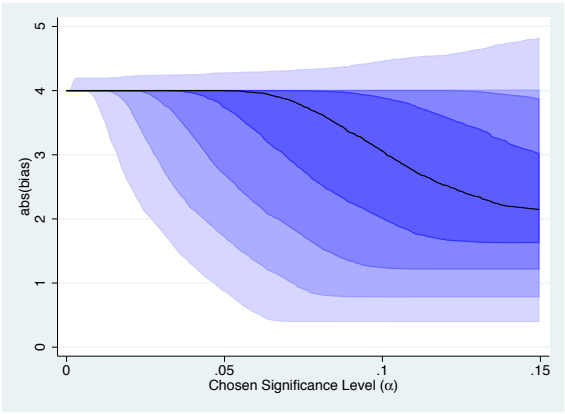


Figure 2: Expected Bias of Forecasts at Different Levels of α

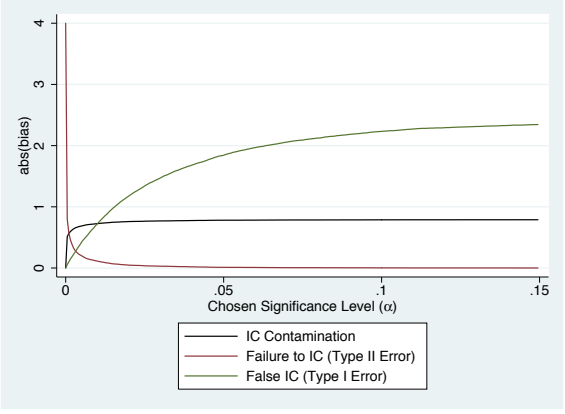
(a) Disaggregate Bias



(b) Aggregate Bias



(c) Disaggregate Bias Decomposition



(d) Aggregate Bias Decomposition

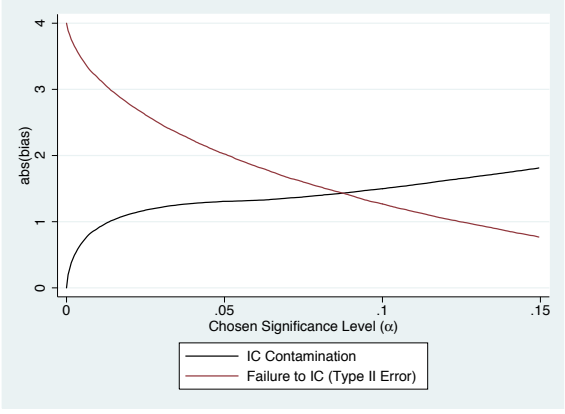


Figure 3: Difference in Median Bias, Aggregate Approach minus Disaggregate Approach

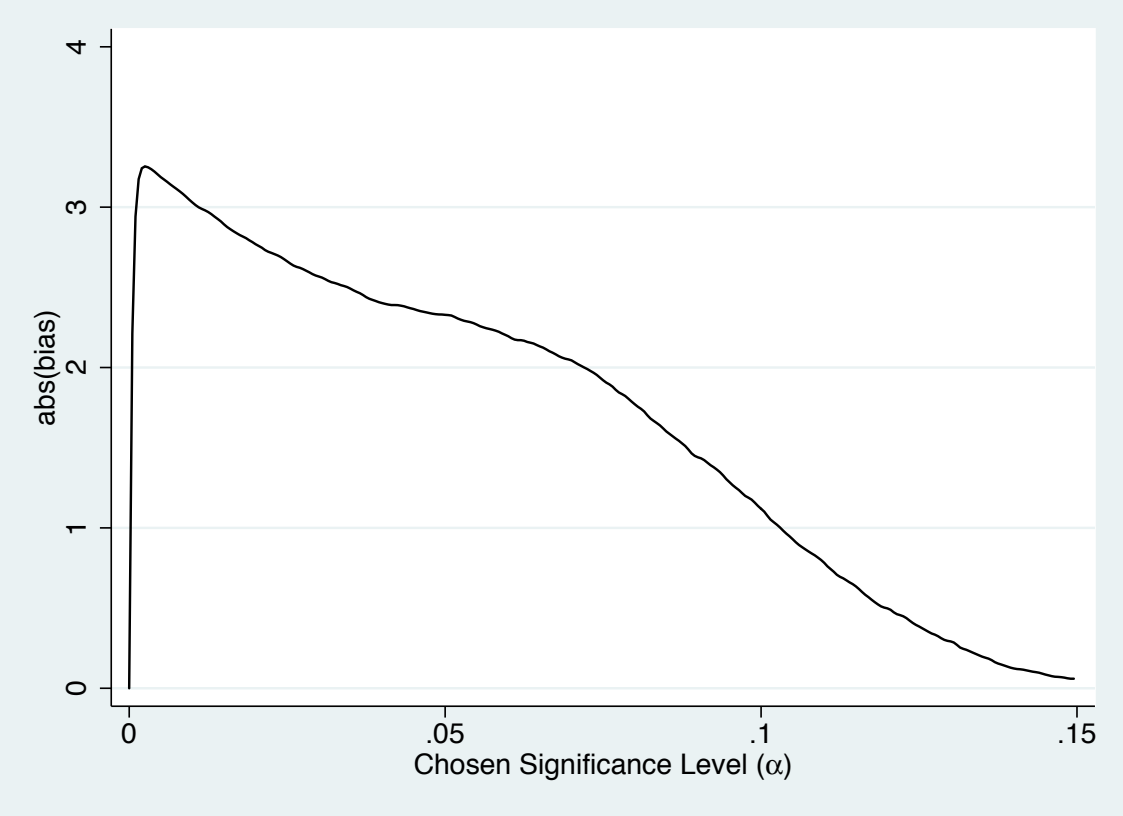


Figure 4: Proven Oil Reserves and Prices, 1986-2012

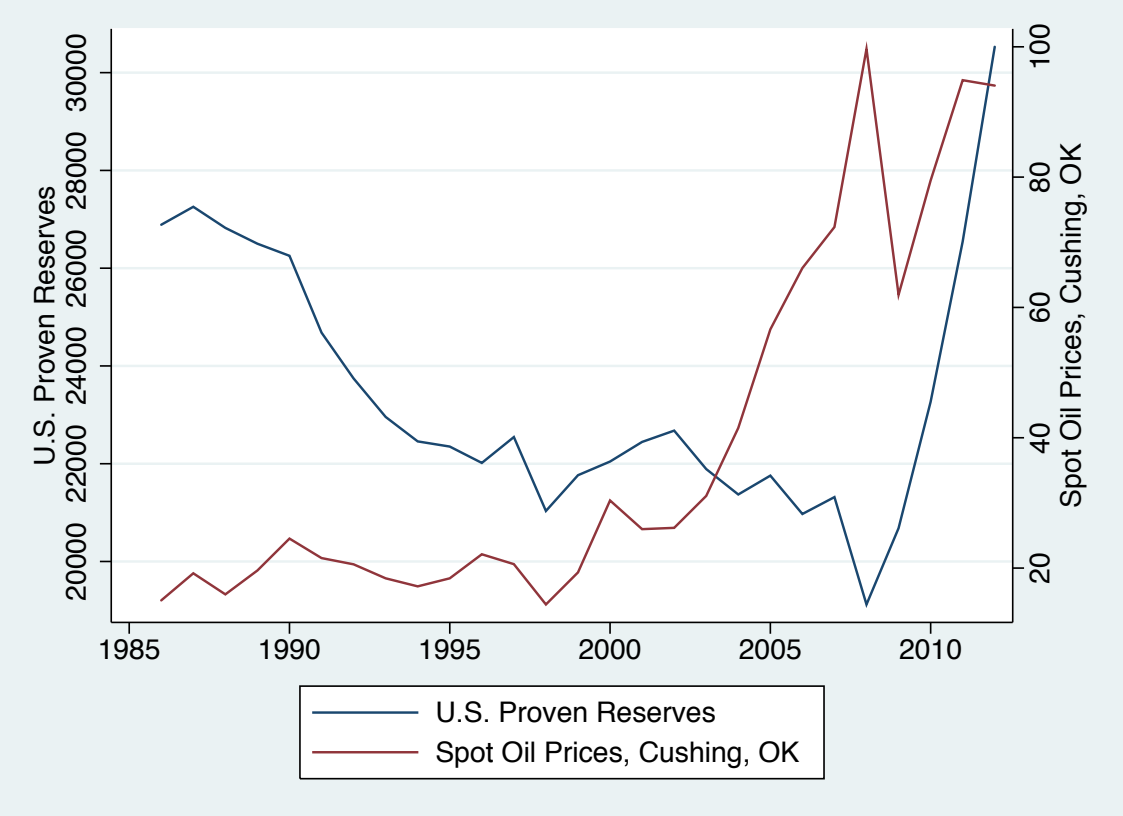
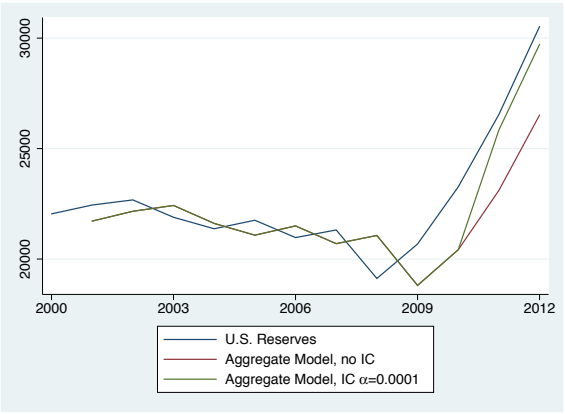
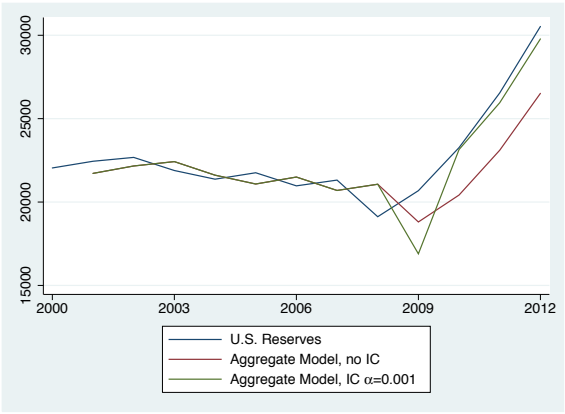


Figure 5: Forecasts using Aggregate-Only Models

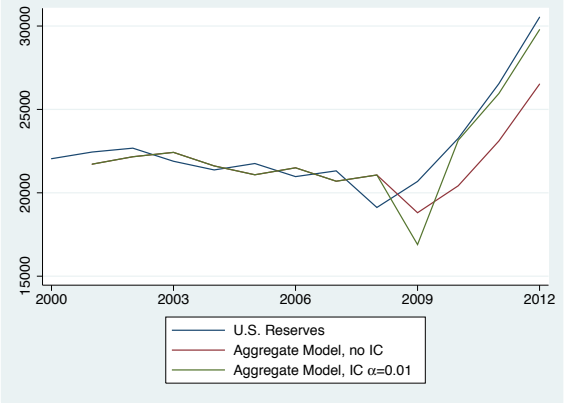
(a) $\alpha = 0.0001$



(b) $\alpha = 0.001$



(c) $\alpha = 0.01$



(d) $\alpha = 0.1$

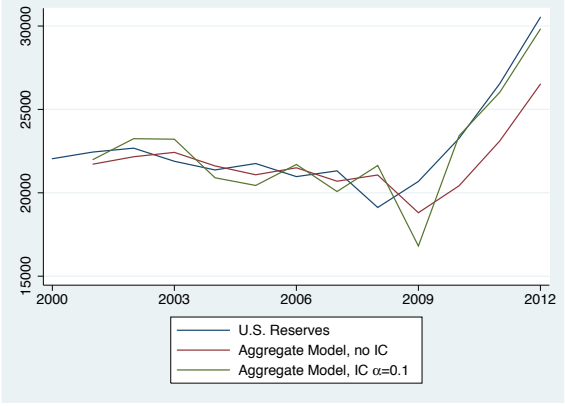
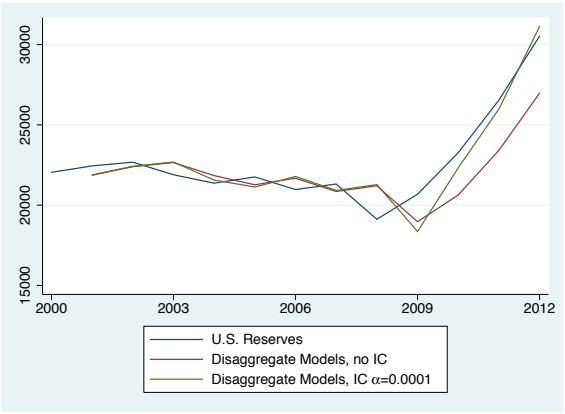
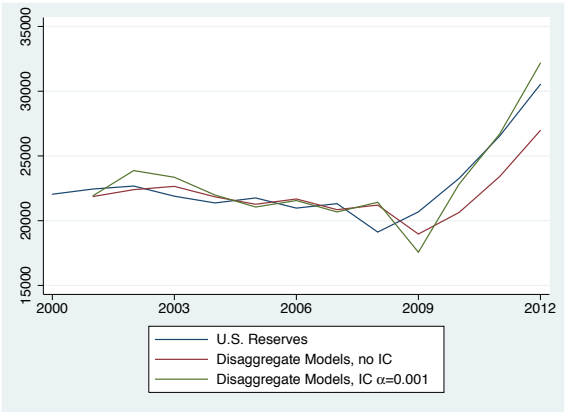


Figure 6: Forecasts using Disaggregate Models

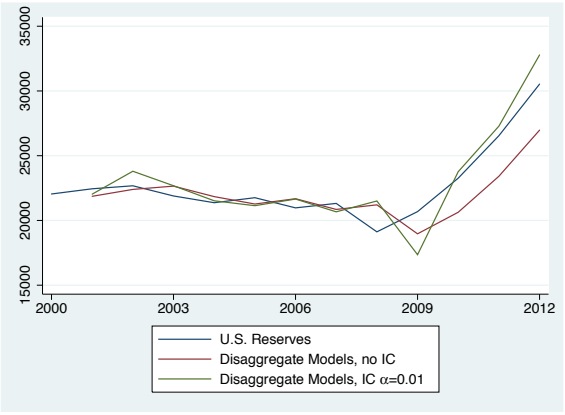
(a) $\alpha = 0.0001$



(b) $\alpha = 0.001$



(c) $\alpha = 0.01$



(d) $\alpha = 0.1$

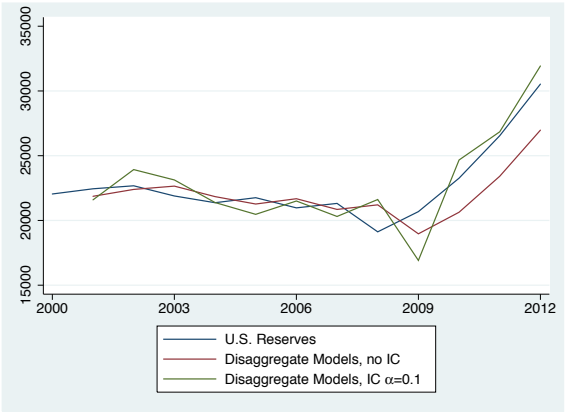


Figure 7: Forecasts using Aggregate Models with Disaggregate Information

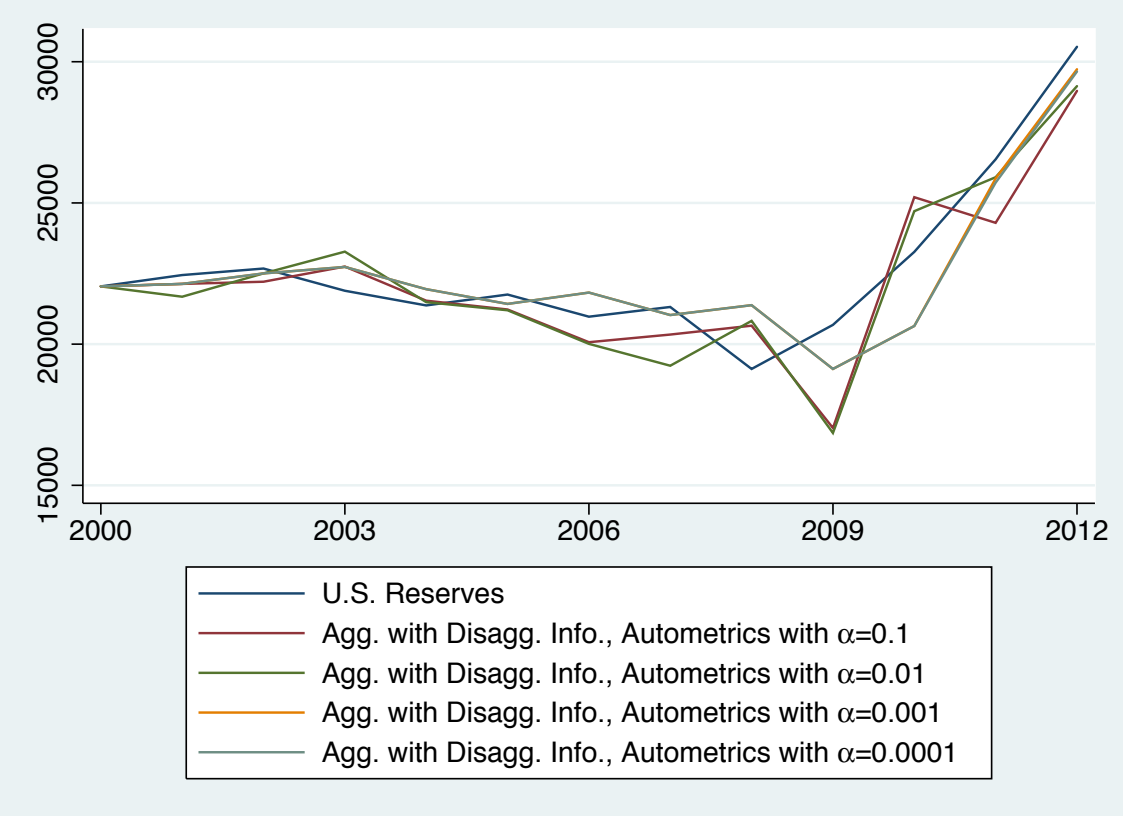


Figure 8: Best Forecasts using Alternative Modeling Approaches

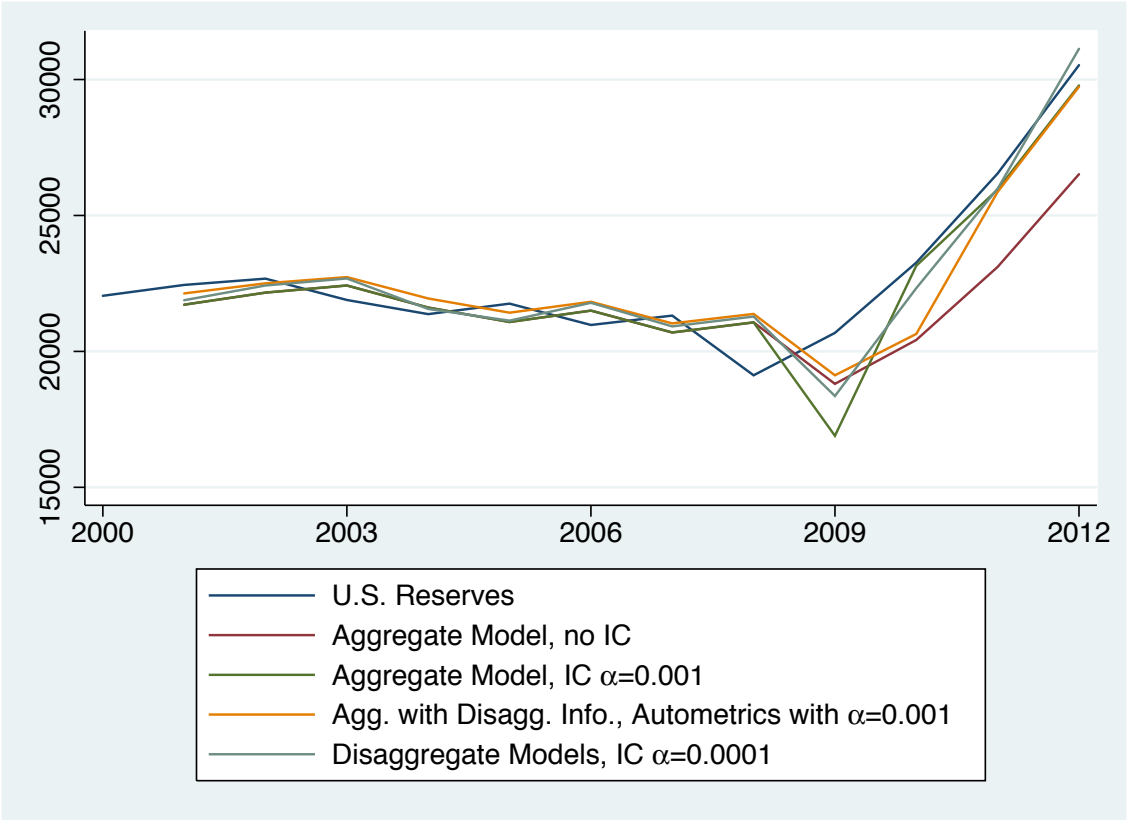


Figure 9: Fraction of Disaggregates with Identified Mean Shifts in Particular Year

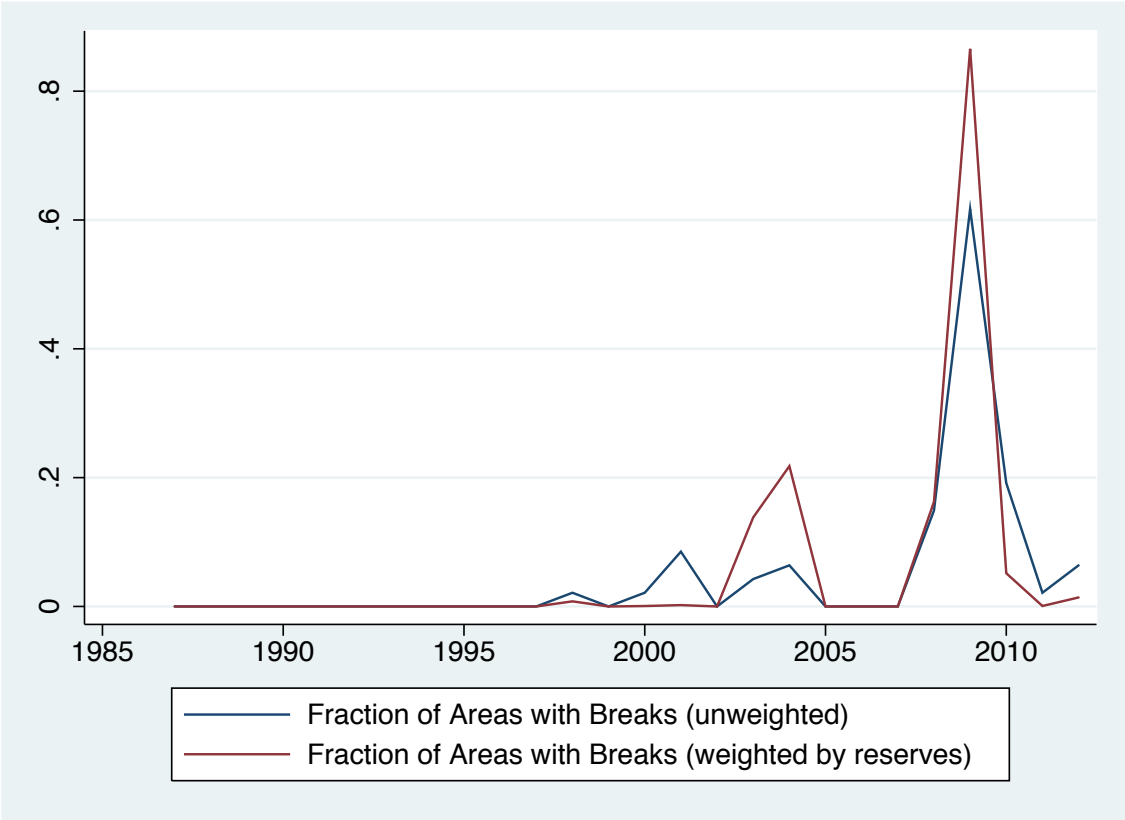


Figure 10: Forecast Combinations: Switching and Average Forecasts