Applied Epidemiological Forecasting with State Space Models

By
Robert A. Yaffee, Ph.D. New York University, New York.
Kent Wagoner, Ph.D. Ithaca College, Ithaca, New York.
Rick Douglass, Ph.D. Montana Technical University, Butte, Mt.
Brian R. Amman, Ph.D. Centers for Disease Control and Prevention, Special Pathogens Branch, Atlanta, Ga.
Thomas Ksiazek, Ph.D. Centers for Disease Control and Prevention, Special Pathogens Branch, Atlanta, Ga.
James N. Mills, Ph.D. Centers for Disease Control and Prevention, Special Pathogens Branch, Atlanta, Ga.
Kostas Nikolopoulos, Manchester Business School, Manchester, U.K.
David Reilly, Automatic Forecasting Systems, Hatboro, Pa,
Sven F. Crone, Lancaster University, Lancaster, U.K.

28th International Symposium on Forecasting
Nice, France
June 25, 2008

Acknowledgments

- *We thank Drs. Pierre Rollin, and C.J. Peters for CDC support of this research. We also thank our field collaborators for assistance, encouragement, and intellectual stimulation: Dr. Terry Yates, Dr. Cheryl Parmenter, Dr. Bob Parmenter, Dr. Chris Hice, Dr. Charlie Calisher, Jeff Root, Jeff Doty, Dr. Ken Abbott, Rachelle Macintosh, Dr. Amy J. Kuenzi, Dr. Andy Hopkins, and Kevin Hughes.*
- *We are indebted to Professors Siem Jan Koopman, David F. Hendry, Hans-Martin Krolzig, Andrew C. Harvey, Neil Shephard, Neil Ericsson, Jurgen Doornik, Ralph Snyder and Rob Hyndman. They have at various times generously given advice concerning structural time series, econometric data-mining, and forecasting issues.*

Disclaimer

- It is not the policy of the CDC to promote any particular commercial software. However, we did use STAMP (Structural Time series Analyzer, Modeler, and Predictor) version 7.0 by Siem Jan Koopman, Andrew Harvey, Jürgen Doornik and Neil Shephard because it models and forecasts nonstationary series with state space models and an earlier version had been used for modeling and hypothesis testing.
- The conclusions and implications are the positions of the authors themselves.

Outline

- Acknowledgements/disclaimer
- Introduction/Background
  - The outbreak
  - The morbidity and mortality
  - The need to forecast
- Research Questions
- Methods
- Results
- Conclusion/Implications
  - Appendix with additional slides
  - Email addresses
  - references

The Outbreak

- In 1993 in the Four Corners region of the Southwestern United States, there was an outbreak of a *frequently fatal* respiratory disease.

- The United States Centers for Disease Control and Prevention (Special Pathogens Branch) was called in to determine the cause of the outbreak.

Morbidity and Mortality

- Early symptoms included fever, headaches, muscle aches, stomach problems, dizziness and chills.
- Later symptoms were coughing, shortness of breath, pectoral tightness, and pulmonary edema.
- Death occurred in about 50% of early cases.

Hantavirus Pulmonary Syndrome (HPS)

- The disease Hantavirus pulmonary syndrome (HPS) was caused by a new hantavirus called Sin Nombre virus (SNV).

- The SNV virus was carried by the deer mouse (*Peromyscus maniculatus*).

The Host Population

1. Over the past 14 years researchers have found that higher host (deer mouse) abundance is associated with an increase in the abundance of infected rodents in an area and subsequently with an increased risk of SNV infection in humans.
2. Forecasting deer mouse populations provides early warning about the ambient epidemiological risk to humans beings.

Forecasting history

- Model Building
  - PcGets was used to select variable for an autoregressive distributed lag model based on an ARIMA model with interventions, presented at CDC research conference in Sevietta, NM.(2002)
  - Using variables selected by PcGets, ad hoc curve fitting models provided good forecasts, as demonstrated at a CDC research conference presentation in Atlanta, Ga. (2003)
  - Variable selection with PcGets version 1. Stamp version 6.3 was used to forecast the nonstationary series CDC Conference in Durango, Col.(2005)
  - Using predictors selected by PcGets version 1.
    - Multiple endogenous lags were used to test density dependence, seasonal component, precipitation, and temperature variables
    - R^2 the best to date
    - Forecasts lacked accuracy
  - Stamp 7.0 was used to forecast the nonstationary MNAtotal series from Montana. For forecasting, we do not model density dependence. The forecasts exhibit much greater accuracy. The report on this analysis follows (2007).

Forecasting History-cont'd

- A Basic structural model was tested: Cyclicity and seasonality and trending slope were found to be not statistically significant.
- Autoregressive lags were tested in the model with Stamp 7.0 (2005-2007).
- A multiple source of error (MSOE) State Space Model was tested (2008)
- Altogether we tested 19 different forecast methods/software.

Forecasting History- 3

- We chose state space methods because the series was nonstationary and an augmented Kalman filter could handle this problem. Yet the state space method was more complicated than many.
- We wanted software and methods that were accurate, simple, and easy to use.
- We chose the Theta method that won the international M-3 competition and those that were relatively automatic.

Forecasting History - 4

- At a CDC Hantavirus research meeting in Durango, Colorado in 2005, Bob Yaffee presented the state space forecasting results.
- Kent Wagoner suggested that we recruit other forecasters, get their forecasts on the same series, and compare the results.
- At that point, Yaffee asked Kostas Nikolopoulos, David Reilly, and Sven F. Crone to participate.
- Kostas presented three versions of the Theta model univariate forecasts.
- Reilly generated two from a causal and a univariate model from Autobox.
- Crone generated univariate feed-forward exhaustive grid search neural network forecasts.
- We compared these according to MAE, MAPE, and MedAPE and ultimately obtained the following results..

Mean Absolute Error

Method	*MAE*	*MAE_rank*
Stamp LLM	18.2729	1
Naïve	18.2778	2
Stamp-LLM+interv	18.9329	3
SES Stata	18.7811	4
Theta AN	19.9575	5
Theta BH	20.0399	6
Forecast Pro	20.0679	7
Damped Trend SPS	20.2755	8
Damped Trend FP	20.6913	9
Theta General	21.9275	10
Theta Opt Wtg	22.7314	11
HW Lin Trend Stata	23.3750	12
HW Add Seas Stata	27.2713	13
SSOE State space	28.7904	14
Autobox-causal	30.8313	15
HW Mult Seas Stata	32.2643	16
Stamp-causal	33.5151	17
Autobox-univariate	33.9284	18
ANN FF	46.2160	19

Mean Absolute Percentage Error

Method	*MAPE*	*MAPE rank*
Stamp LLM	30.8445	1
Damped Trend FP	31.4732	2
Damped Trend SPSS	31.5207	3
Naïve	31.5959	4
Stamp LLM + interv	31.1105	5
SES Stata	32.3526	6
Theta AN	32.5539	7
Forecast Pro	32.6062	8
Theta BH	32.7424	9
Theta General	39.8989	10
SSOE State space	40.6046	11
Theta Opt Wtg	41.4875	12
Stamp-causal	42.1822	13
HW Lin Trend Stata	43.3061	14
HW add Seas Stata	43.5801	15
HW mult Seas Stata	45.6779	16
Autbox-causal	49.2464	17
Autobox-univariate	50.4155	18
ANN FF	63.6246	19

Median Absolute Percentage Error

Method	*MedAPE*	*Med_rank*
Stamp LLM	18.5468	1
Autobox-causal	19.0708	2
Theta AN	19.0981	3
SES Stata	20.0127	4
Theta BH	20.1417	5
Stamp LLM + interv	20.1553	6
Theta General	20.4098	7
Damped Trend SPS	22.0576	8
Naïve	22.1014	9
Stamp-causal	22.4163	10
Forecast Pro	22.7373	11
Theta Opt Wtg	22.8761	12
Damped Trend FP	23.0143	13
SSOE State space	26.6849	14
HW Lin Trend Stata	32.0758	15
Autobox-univariate	34.0471	16
HW Add Seas Stata	38.9371	17
HW Mult Seas Stata	42.3424	18
ANN FF	53.0188	19

- Because the local level (MSOE) state space model performed so well according to these criteria, we wondered why it outperformed the causal model.

- What's novel about this inquiry?
  - This study compares causal and univariate models.
  - It entails a rolling origin forecast evaluation.
  - This study includes the use of state space methods in such a comparison.
  - It even includes a single source of error (SSOE) state space method.
  - Recognizing that these methods may be software dependent, we rely on the default parameters in the software to provide the results.
  - We compare a multiple source of error state space model to a single source of error state space model.

Research Questions

1. Can we forecast the abundance of the deer mouse population with readily available weather data?

2. Can state space methods forecast the MNAtotal series with readily available weather predictors and interventions?

3. Can state space methods forecast the MNAtotal series with a local level model and interventions?

4. Can we generate a good a univariate forecast of MNAtotal without the weather variables ?

5. Can Stamp forecast the MNAtotal from a univariate local level model (random walk plus noise)?

6. Which forecast is most accurate?

Methods

- State space models consist of a measurement equation, a transition equation, and conditions of initialization.
- We test components for significance and find that cyclicality and seasonality are not significant measurement model components and therefore exclude them from our models.
- We find three measurement models to be of promise with an augmented Kalman filter.
  - *Local Level Model with dynamic parameters and interventions (level shifts and additive outliers).*
  - *Local Level Model with interventions*
  - *Local Level Model without interventions*
- Forecast Protocol

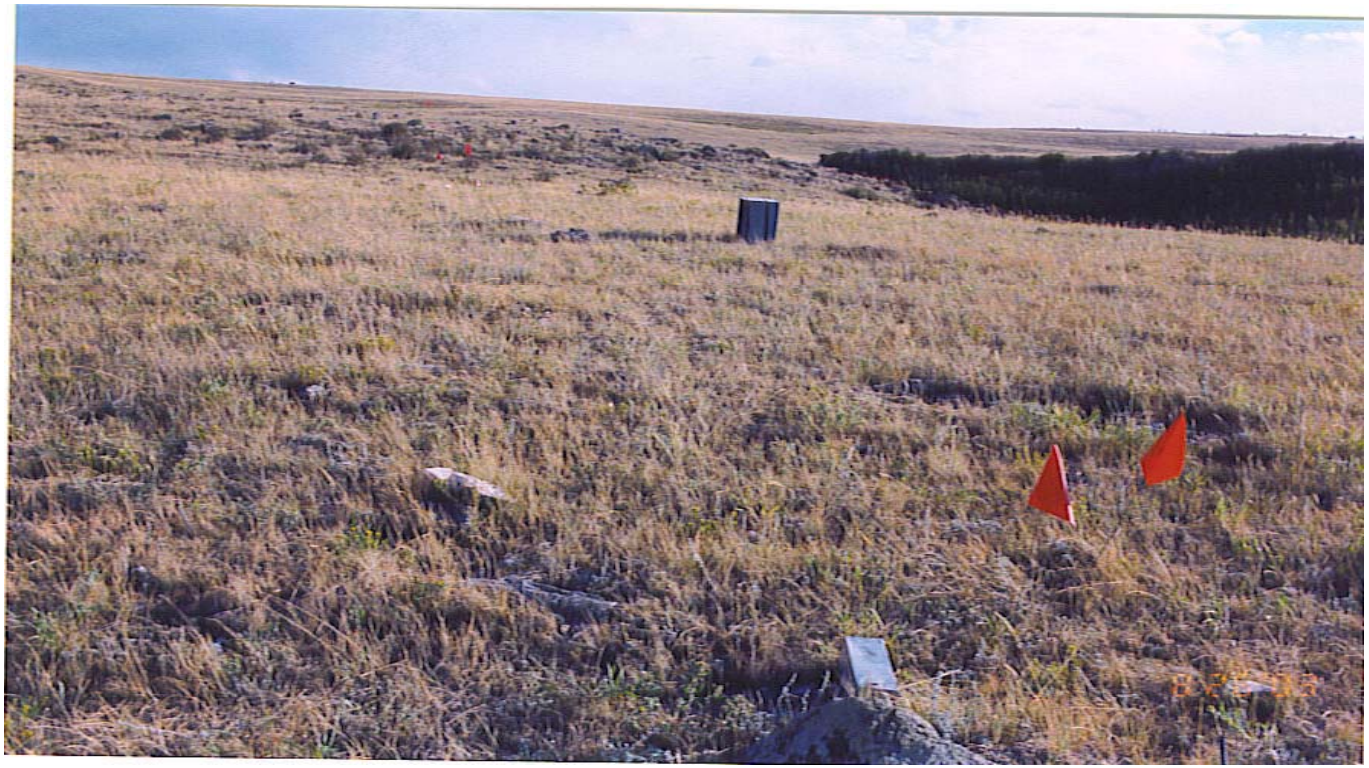
Data Collection

- *Minimum Number Alive (MNA)*
  - *MNA is a measure of population abundance (Chitty and Phipps, 1966)*
  - *MNA is based on capture-recapture method*
  - *100 Traps are set along a grid within 1 hectare*
  - *Captured mice are tagged and released*
  - *MNA(X)=number captured in month X+ those not captured in month X but captured in at least 1 previous month and at least 1 subsequent month.*

Trapping Grids in Central Montana

Site 10

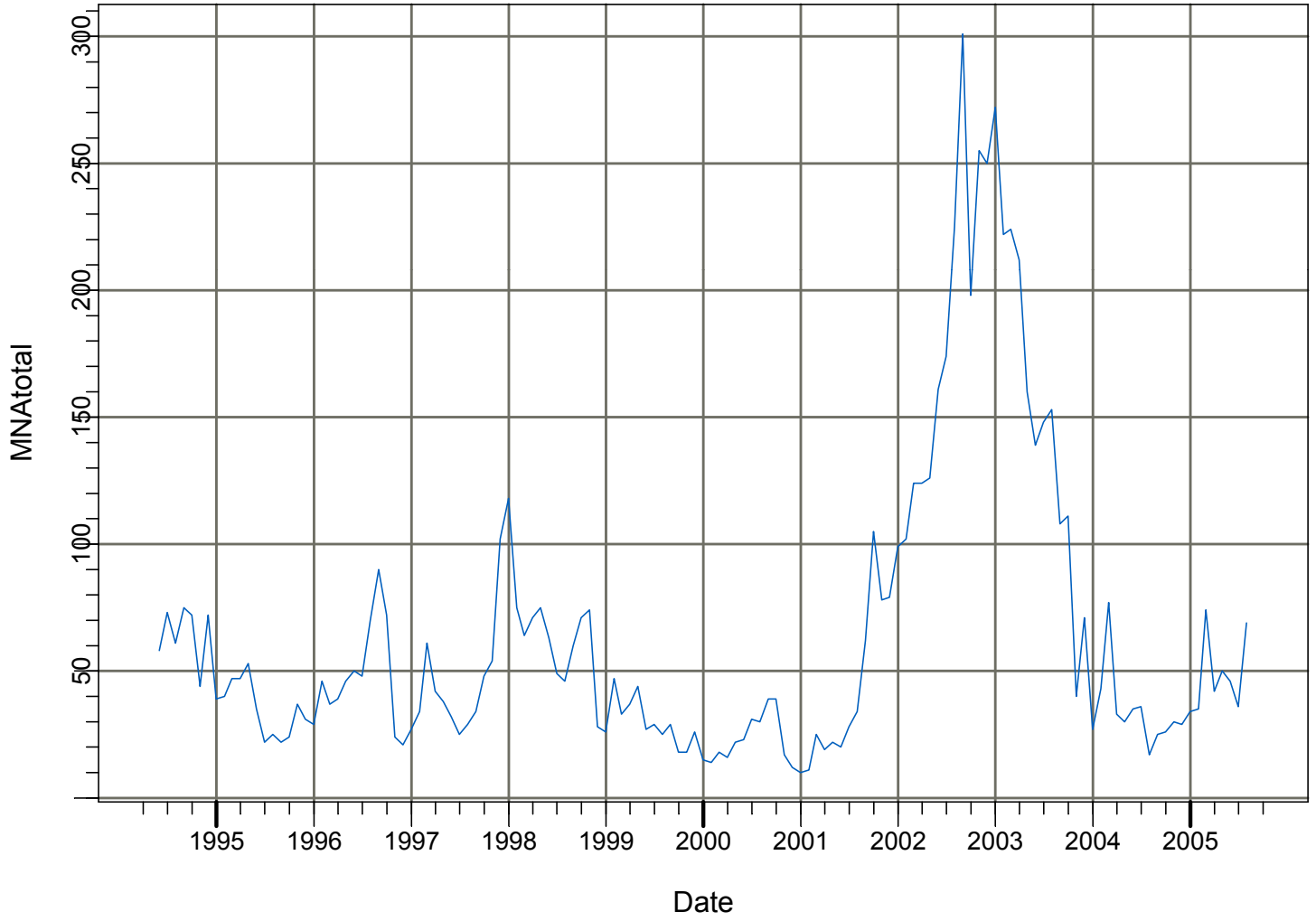
Site 11



Site 12

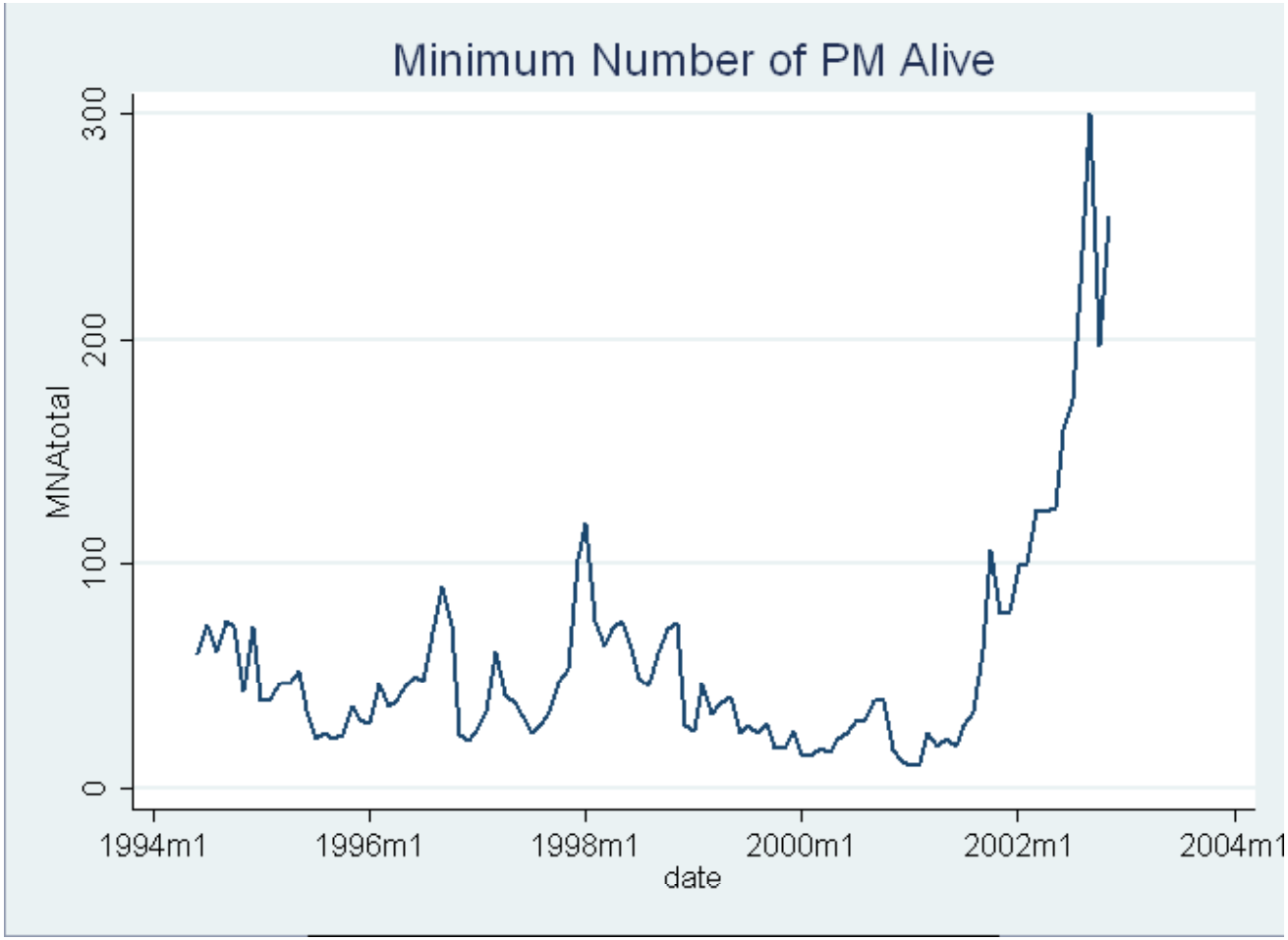
The MNAtotal Series

Abundance of *Peromyscus maniculatus* (deer mouse) in the Montana Cascade



The MNAtotal series

when we began forecasting



Characteristics of the MNA$_{total}$ Series

- MNAtotal is nonstationary. The series moments change. They are best estimated by changing parameters or local trends.
- Depending on the model this series contains at least 3 outliers and 17 level shifts
- The polynomial lags have unstable roots.
- The number of significant endogenous lags varies with the length of the series.
- Numerous end-effects or sudden changes at the end of the series from which the forecast is generated. These effects bias the forecast and impair forecast accuracy.

Chronology of Structural Breaks

	year	month	levelshift	pulse
1	1995	6	levelshift	
2	1997	12	levelshift	
3	1998	2	levelshift	
4	1998	12	levelshift	
5	2001	9	levelshift	
6	2001	10	levelshift	
7	2002	3	levelshift	
8	2002	6	levelshift	
9	2002	8	levelshift	
10	2002	9		outlier
11	2002	11	levelshift	
12	2003	2	levelshift	
13	2003	5	levelshift	
14	2003	9	levelshift	
15	2003	11	levelshift	
16	2003	11		outlier
17	2004	1	levelshift	
18	2004	3	levelshift	
19	2004	4	levelshift	
20	2005	3		outlier

MNA_{total} Structural Breaks: Their Significance, Direction, and Magnitude

mnatotal	Coef.	Semi-robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
mnatotal						
p_2002m9	89.77985	7.963498	11.27	0.000	74.17168	105.388
p_2003m11	-33.77297	3.961494	-8.53	0.000	-41.53736	-26.00858
p_2005m3	35.73247	2.646475	13.50	0.000	30.54547	40.91946
lev_2002m11	52.70329	6.980822	7.55	0.000	39.02113	66.38545
lev_2003m5	-59.11839	7.374715	-8.02	0.000	-73.57257	-44.66422
lev_2002m8	48.62701	8.524699	5.70	0.000	31.9189	65.33511
lev_2003m2	-49.15668	6.046551	-8.13	0.000	-61.0077	-37.30566
lev_2001m10	41.18169	6.278305	6.56	0.000	28.87644	53.48694
lev_2004m1	-39.23094	6.963248	-5.63	0.000	-52.87865	-25.58322
lev_2003m9	-43.91622	4.314937	-10.18	0.000	-52.37334	-35.4591
lev_2002m6	38.26475	5.004345	7.65	0.000	28.45641	48.07308
lev_1998m12	-39.92056	4.421628	-9.03	0.000	-48.58679	-31.25432
lev_1997m12	61.34417	8.704992	7.05	0.000	44.2827	78.40564
lev_1998m2	-37.09875	7.022756	-5.28	0.000	-50.8631	-23.3344
lev_2001m9	29.62409	4.216216	7.03	0.000	21.36046	37.88772
lev_2004m4	-38.68764	4.914898	-7.87	0.000	-48.32066	-29.05462
lev_2004m3	39.82155	5.178501	7.69	0.000	29.67187	49.97123
lev_2002m3	27.40037	5.972633	4.59	0.000	15.69423	39.10652
lev_2003m11	-35.18214	4.877162	-7.21	0.000	-44.7412	-25.62308
lev_1995m6	-16.5651	6.834973	-2.42	0.015	-29.9614	-3.168797
_cons	57.15984	6.069309	9.42	0.000	45.26421	69.05547
ARMA						
ar						
L1.	.5131535	.1208405	4.25	0.000	.2763105	.7499966
SIGMA2						
_cons	103.6064	17.11018	6.06	0.000	70.07108	137.1417

Forecast Protocol:

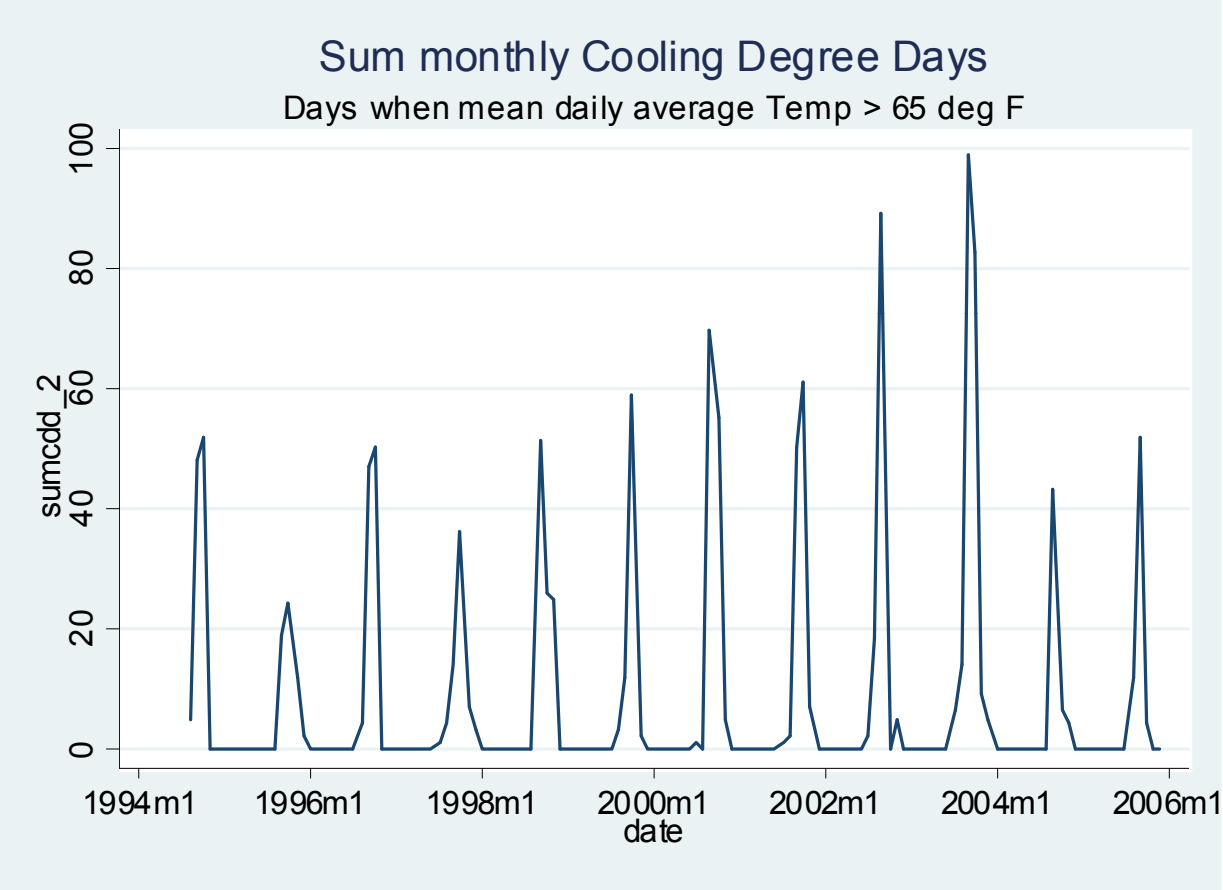
Rolling origin *ex ante* 3 month forecast

- The purpose was to average out unusual end effects biasing of forecasts.
- Time Series Length. June 1994 – November 2005, inclusive [138 observations].
- First Point of Forecast Origin: We begin forecasting three months ahead in November 2002 [observation 102].
- Rolling Origin Forecast: The origin of the forecast is rolled ahead one (three month) season for each forecast [102, 105, …, 135]. The next forecast extends over the next season. Seasons are defined as:
  - Winter: Dec through Feb
  - Spring: March through May
  - Summer: June through Aug
  - Fall: Sep through Nov.
- The process is reiterated.
- In sum, 12 three month *ex ante* forecasts were generated from datasets up to the point of forecast origin.
- We estimated the average error over the three month period.

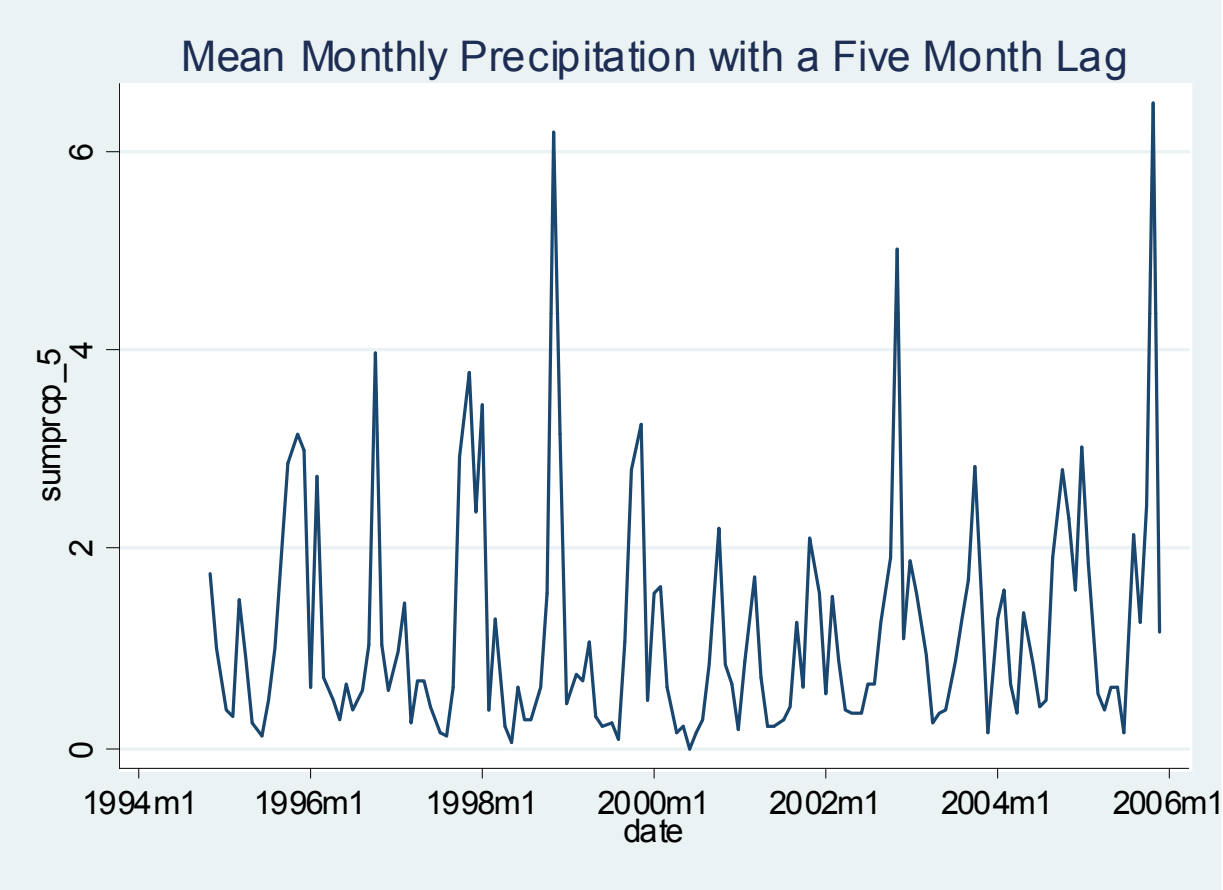
The Conditional Model: Significant Weather Predictors

- Significant
  - Sum of cooling degree days (lag 2) mostly
  - Sum of monthly precipitation (lag 5) a few times
- Not significant
  - Sum of monthly snow (inches)
  - maximum monthly temperature (degrees F)
  - Minimum monthly temperature (degrees F)
  - Average monthly temperature (degrees F)
  - Sum of monthly heating degree days
  - Mean Monthly Temperature (degrees F)

Sum of Cooling Degree Days with a two month lag



Sum of Monthly Precipitation with a Five Month Lag



Impact of these predictors?

- We found that warmer temperatures could lead to greater mouse population. But that the effect was not a dominant one. Rather, it was a minimal one.
  - *Sum of cooling degree days at lag 2 is a stable predictor. It is significant but not large. It only accounts for a 2.718^.004 =1.004 increase in MNAtotal with a 1 degree Fahrenheit rise in temperature. This suggests that warmer weather is associated with an increase in the abundance of the mouse population. The lag 2 may arise from the relatively cold weather in Montana.*
  - *Mean minimum monthly temperature suggests that as this goes up so does the prevalence of the deer mice. Both of these indicators suggest that warmer temperatures may be important. This has implications for global warming and the diffusion of this disease.*
  - *The impact of the five month lag of monthly precipitation was even smaller.*

Distribution of Level Shifts in Stamp Causal Models

numlevshift

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3	8.3	8.3	8.3
	1	6	16.7	16.7	25.0
	2	3	8.3	8.3	33.3
	3	12	33.3	33.3	66.7
	4	12	33.3	33.3	100.0
	Total	36	100.0	100.0	

The number of level shifts also varies with the data range.

Distribution of Additive Outliers in Stamp Causal Models

numaddout

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6	16.7	16.7	16.7
	1	6	16.7	16.7	33.3
	2	3	8.3	8.3	41.7
	3	6	16.7	16.7	58.3
	4	3	8.3	8.3	66.7
	5	3	8.3	8.3	75.0
	6	3	8.3	8.3	83.3
	7	3	8.3	8.3	91.7
	8	3	8.3	8.3	100.0
	Total	36	100.0	100.0	

Similarly, the number of additive outliers may vary with local level of the data.

Why State Space Models should be able to Forecast this series

- The MNA_{total} series is nonstationary.

- State Space models with augmented Kalman filters can forecast nonstationary series. They use an extended random normal vectors [and matrices] that define the means, variances, and covariances and other information needed to define the state of a system at a given point in time (Snyder and Forbes, 2002, p.3).

- There are many level shifts in the MNAtotal series, which may be able to be modeled and forecast by a method that can model and forecast a local level model (Muth, 1960).

The State Space Method

- Initial condition
  - Takes an initial estimate of the mean and the variance
- Transition equation sequentially updates the estimate
  - Kalman filter computes the mean and variance of the state vector, given the prior information and sequentially updates in with a first-order autoregression plus a regression on the innovation.
- Measurement equation
  - A linear combination of components in the model from which a measurement of error is obtained. The state is estimated and the error is minimized.
- Correction is performed by minimizing the predictive error variance.
  - The process reiterates until convergence is obtained.
- Forecasting is performed with the transition equation

The General Measurement Equation

(Koopman, Harvey, Doornik, and Shepard, 2006, p.143)

$$y_t = \mu_t + \beta_t + \gamma_t + \psi_t + \sum_{\tau=1}^{p} \phi_\tau y_{t-\tau} + \sum_{i=1}^{k}\sum_{\tau=0}^{q} \Delta_{it} x_{i,t-\tau} + \sum_{j=1}^{h} \lambda_j I_j + \varepsilon_t$$

where $y_t = \ln(MNAtotal)$

$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t =$ *trend for* $t = 1, \ldots, n$

$\beta_t = \beta_{t-1} + \varsigma_t =$ *slope*

$\gamma_t =$ *seasonality*

$\psi_t =$ *cyclical component*

$\phi_t =$ *autoregressive parameters*

$\sum_{i=1}^{k}\sum_{\tau=0}^{q} \Delta_{it} x_{i,t-\tau} =$ *dynamic (time varying) parameter estimates* $_{\tau = \text{time lag}}$

$\Delta_{it} =$ *parameter estimates (not differences)*

$\sum_{j=1}^{h} \lambda_j I_j =$ *Interventions(level shifts, slope shifts, outliers)*

$\varepsilon_t =$ *observation error*

The Measurement (Observation) Equation

$$y_t = Z_t \alpha_t + \varepsilon_t$$

where

$Z_t = selection\ \ matrix\ of\ \ factor\ \ loadings$

$\alpha_t = state\ vector$

$\varepsilon_t = observation\ error\ \ matrix$

$\varepsilon_t \sim NID(0, H_t)$

$H_t = observation\ error\ \ variance\ \ matrix$

Component Loading into State Vector

- Components load into the State vector for a model with a local level, and a monthly seasonal component (11 dummy variables):

$$y_t = (1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0)\alpha_t + (1\,1)\varepsilon_t$$

where

$$\alpha_t = \begin{pmatrix} u_t \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \\ \gamma_{4,t} \\ \gamma_{5,t} \\ \gamma_{6,t} \\ \gamma_{7,t} \\ \gamma_{8,t} \\ \gamma_{9,t} \\ \gamma_{10,t} \\ \gamma_{11,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \varepsilon_t$$

$$\varepsilon_t = \begin{pmatrix} \eta_t \\ \omega_t \end{pmatrix}$$

Transition Equation

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t$$

where

α_t = *the unobserved state vector*

T_t = *state projection matrix*

R_t = *innovation selection matrix*

η_t = *innovation vector describing change in* α_t *for local level model from time* t *to time* $t+1$

$\eta_t \sim NID(0, Q_t)$

Q_t = *evolution disturbance variance matrix for* α_t

η_t *and* ε_t *are assumed to be independent*

Initial conditions

$$\alpha_0 = N(\alpha_0, P_0)$$

where

$$\alpha_0 = initial\ mean\ of\ state\ vector$$

$$P_0 = initial\ state\ estimation\ error\ variance$$

Estimation

One solves for the component parameter estimates by minimizing via maximum likelihood the predictive error variance, F_t *:*

where

$$F_t = var(v_t)$$
$$= E[(y_t - E(y_t | Y_{t-1})(y_t - E(y_t | Y_{t-1})'$$
$$= E(\alpha_t - \bar{\alpha})(\alpha_t - \bar{\alpha})'$$
$$= E(y_t - Z_t\alpha_t)(y_t - Z_t\alpha_t)'$$
$$= Z_t P_t Z_t' + H_t$$

and

$$P_t = E(Var(\alpha_t | Y_{t-1})) = F_t - \sigma_\varepsilon^2$$

σ_ε^2 = *measurement (observation) error variance*

(Durbin and Koopman,(2000),12,26 - 27,66).

Maximum Likelihood

- The likelihood maximized is a function of the prediction error variance by a Broyden, Fletcher, Goldfarb and Shanno algorithm:

$$LL = -\frac{n}{2}\log 2\pi - \frac{1}{2}\left(\sum_{t=1}^{t}\log|F_t| + \sum_{t=1}^{t}F_t^{-1}v_t^2\right)$$

where

$LL = log\ \ likelihood$

$v_t = innovation$

$F_t = var(v_t)$

(Durbin and Koopman, 2000, 138)

Updating (correcting) and Forecasting Equations

$\alpha_{t+1} = T_t \alpha_t + K_t v_t \quad for\ mean$

$P_{t+1} = P_t(1-K) + \sigma_\eta^2 \ for\ the\ variance$

where

$T_t = an\ m\ x\ m\ state\ projection\ matrix$

$\alpha_t = state\ vector$

$a_t = \bar{\alpha} = \ average\ state\ vector$

$v_t = y_t - a_t \ (innovation)$

$P_t = Var(\alpha_t) = state\ variance$

$F_t = predictive\ error\ variance$

$K_t \ = \ P_t/F_t = Kalman\ gain \left(\frac{state\ evolutionary\ variance}{predictive\ error\ variance} \right)$

$\sigma_\eta^2 = Q_t = innovation\ variance\ matrix$

(Durbin and Koopman, 2000, 12,66)

Kalman Filtering

- Sequential updating: The Kalman filter updates the new variance estimate from the current and previous variances. The updating is performed by a weighted average. The weights are formulated from the precisions (inverse of the variances) of the likelihood and the prior probability distribution. It also uses this to update the mean
- Diffuse prior: When nothing is known about the previous variance, an noninformative prior is assumed. The inverse of a very large (diffuse) variance accords it small precision and little weight.

Bayesian sequential updating

Guido Imbens(2007) "What's New in Econometrics Lecture 7 on Bayesian Inference National Bureau of Economic Research Summer Institute, 2007, p. 6.

Suppose the Model $\sim N(\mu, \sigma^2)$

and the prior distribution for $\mu \sim N(\mu_0, \tau^2)$.

Sequential updating takes by:

$$E(\mu \mid x) = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \quad \textit{and}$$

$$V(\mu \mid x)^{-1} = \textit{Precision} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$$

$$\textit{Hence: Posterior distribution} \sim N\left(\frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \right)$$

Augmented Kalman Filter

- Following earlier works of Ansley and Kohn (1985), DeJong(1991) and Rosenberg (1993), Harvey(1993, 138) and Durbin and Koopman (2000,113-117) formulate an augmented Kalman filter to permit the analysis of nonstationary series:
- They augment the state vector in a partition of stationary and nonstationary components.
- The partitioning allows the assignment of a diffuse prior to the nonstationary components.
- Because weighting in the sequential updating process is done in accordance with the precision or inverted variance, when little is known about the nonstationary elements, a diffuse prior (with an arbitrarily very large variance)—P_∞—is used for weighting. When a singular matrix is inverted, generalized inverses are used.
- After a few extra iterations, the process converges to a proper solution.

We estimate 3 Cases with this augmented Kalman filter

- Case 1: Local level model, time varying parameters, and interventions.

- Case 2: Local level model plus interventions. Taylor's approach.

- Case 3: Local level model: The baseline.

Case 1: Local Level Model + time varying parameters + interventions

$$y_t = \mu_t + \sum_{i=1}^{k}\sum_{\tau=0}^{q}\Delta_{it}x_{i,t-\tau} + \sum_{j=1}^{h}\lambda_j I_j + \varepsilon_t$$

$$where\ y_t = \ln(MNAtotal)$$

$$\mu_t = \mu_{t-1} + \eta_t = \ level\ for\ t = 1,...,n$$

$$\sum_{i=1}^{k}\sum_{\tau=0}^{q}\Delta_{it}x_{i,t-\tau} = time\ varying\ \ parameter\ estimates$$

$$\sum_{j=1}^{h}\lambda_j I_j = Interventions(level\ shifts,\ slope\ shifts, outliers)$$

$$\varepsilon_t = \ error\ or\ disturbance$$

Case 2: Local Level Model + Interventions Measurement Equation

$$y_t = \mu_t + \sum_{i=1}^{k}\sum_{\tau=0}^{q} \Delta_{it} I_{t-\tau} + \varepsilon_t$$

where $y_t = \ln(MNAtotal)$

$\mu_t = level \quad for\ t = 1,...,n$

$\Delta_{it} =$ *parameter estimates* (*not differences here*)

$I_{t-\tau} = Intervention$ (*outlier*, *level shift*, *slope shift*)

$\tau =$ *time lag*

$\varepsilon_t =$ *error or disturbance*

if $\mu_t = \alpha_t$ *where* $\alpha_t =$ *random walk*

where all random variables are normally distributed

and ε_t *has constant variance.*

Case 3: Local Level Model
Measurement Equation

$$y_t = \mu_t + \varepsilon_t$$

where $y_t = \ln(MNAtotal)$

$\mu_t = level \quad for \; t = 1, \ldots, n$

$\varepsilon_t = \; error \; or \; disturbance$

if $\mu_t = \alpha_t$ *where* $\alpha_t = random \; walk$

where all random variables are normally distributed and ε_t *has constant variance.*

State Space stacked matrix Formulation

- Local level model (random walk plus noise) (Zivot, Wang and Koopman (2004), 287).

$$If\ \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t = iid\ N(0,\sigma^2_{\eta_t})\ transition\ equation$$

$$y_t = Z_t\alpha_t + \varepsilon_t, \qquad \varepsilon_t = iid\ N(0,\sigma^2_{\varepsilon_t})\ measurement\ equation,$$

then

$$\begin{pmatrix}\alpha_{t+1}\\ y_t\end{pmatrix} = \begin{pmatrix}T_t\\ Z_t\end{pmatrix}(\alpha_t) + \begin{pmatrix}R_t\eta_t\\ \varepsilon_t\end{pmatrix}$$

where

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t = iid\ N(0,\sigma^2_{\varepsilon_t})$$

$$\alpha_1 \sim N(\alpha_1, P_1)$$

$$NB: In\ a\ local\ level\ Model: T_t = I,\ so$$

$$\alpha_{t+1} = \alpha_t + R_t\eta_t$$

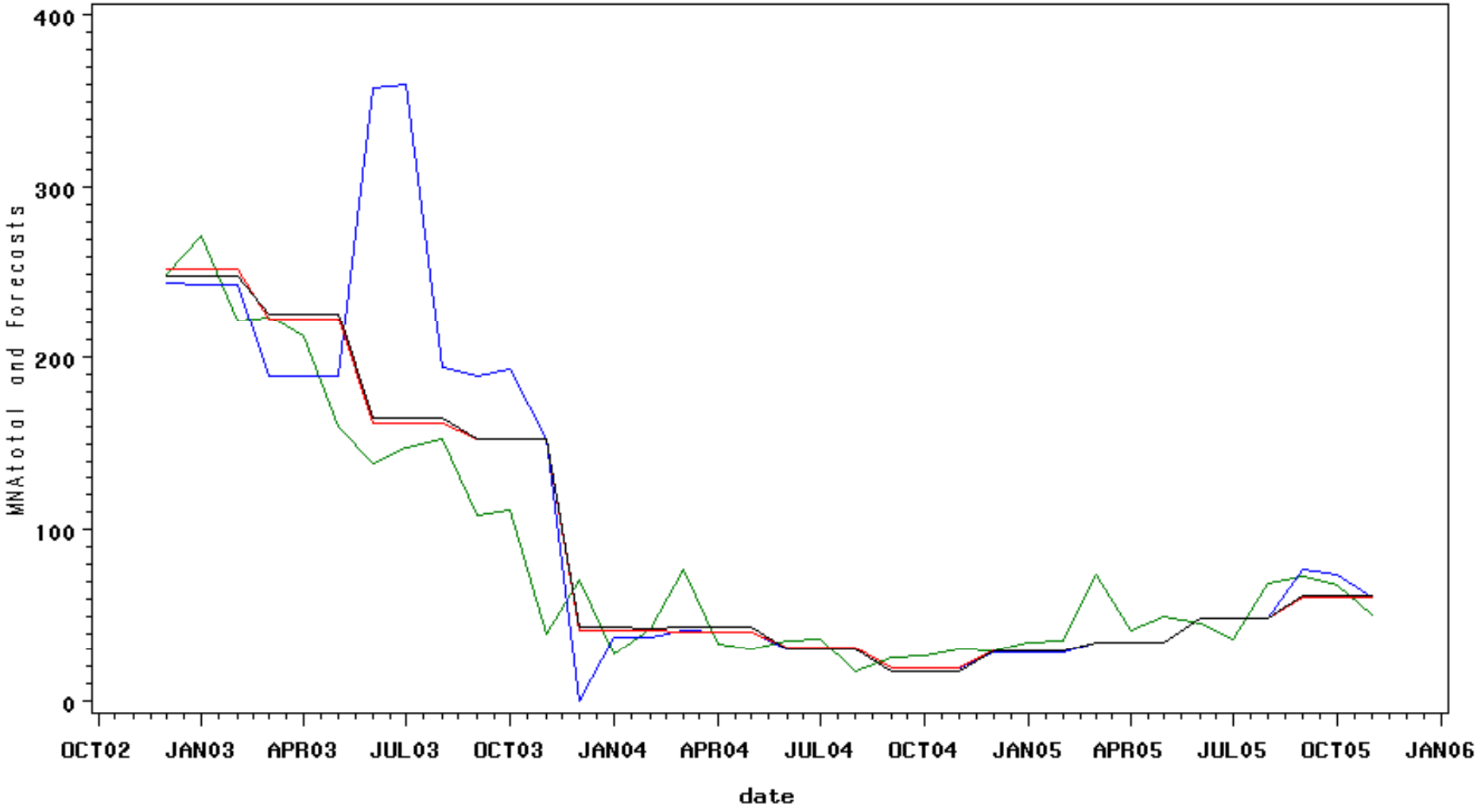
The state vector is furthermore partitioned into stationary and nonstationary components and the diffuse prior is applied to the latter.

Local level Model with Intervention Analysis

- Significant Interventions may also vary in a local level model.
  - Outliers
  - Level shifts

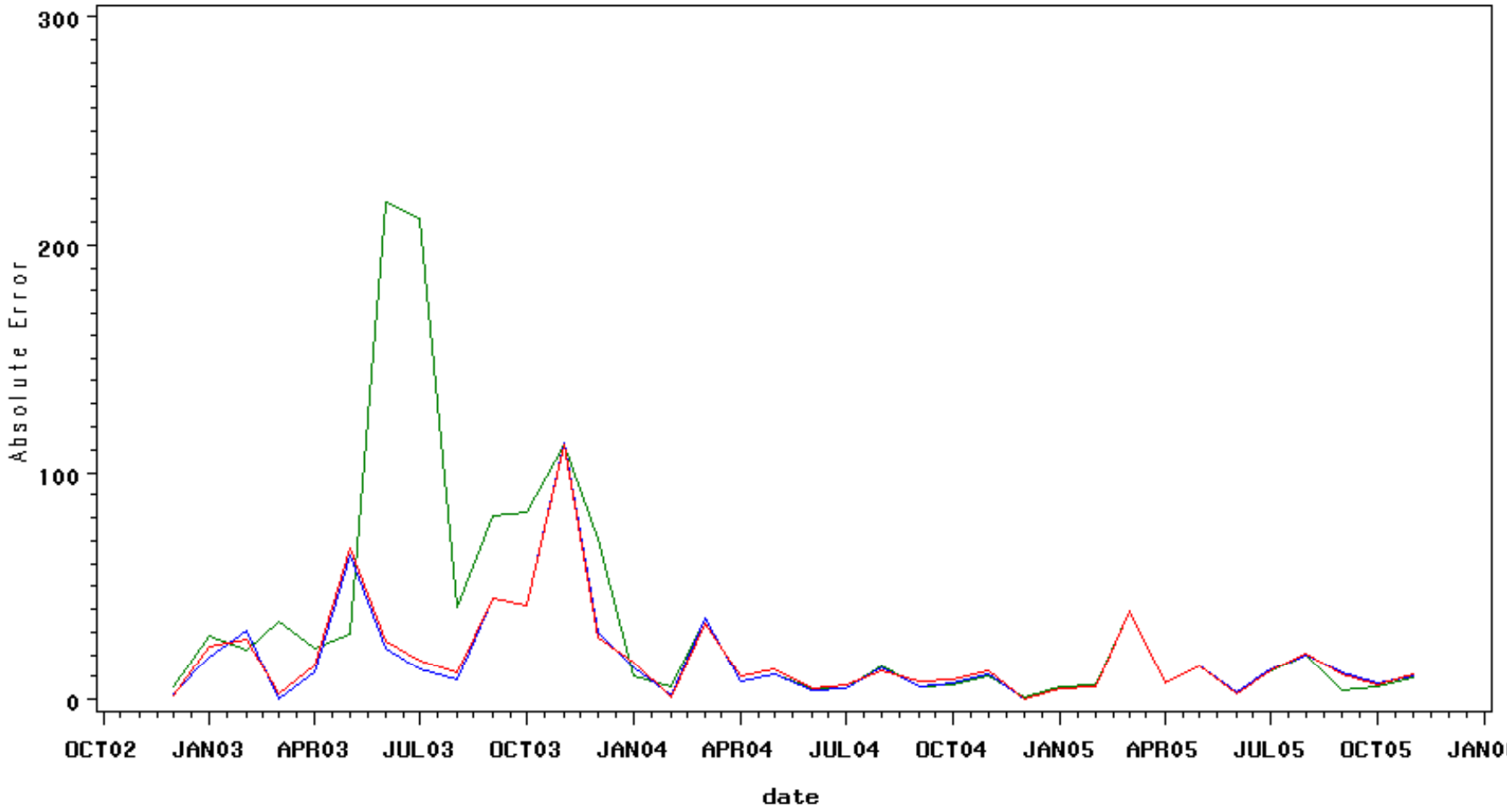
Forecasts and MNAtotal for Three Stamp models

MNAtotal=green Forecast stamp causal = blue
Stamp LLM forecast= red Stamp LLM+interv forecast=black



Absolute error for Three Stamp Forecasts

Absolute error stamp causal = green
Absolute error Stamp LLM forecast= blue
Absolute error Stamp LLM+interv forecast=red



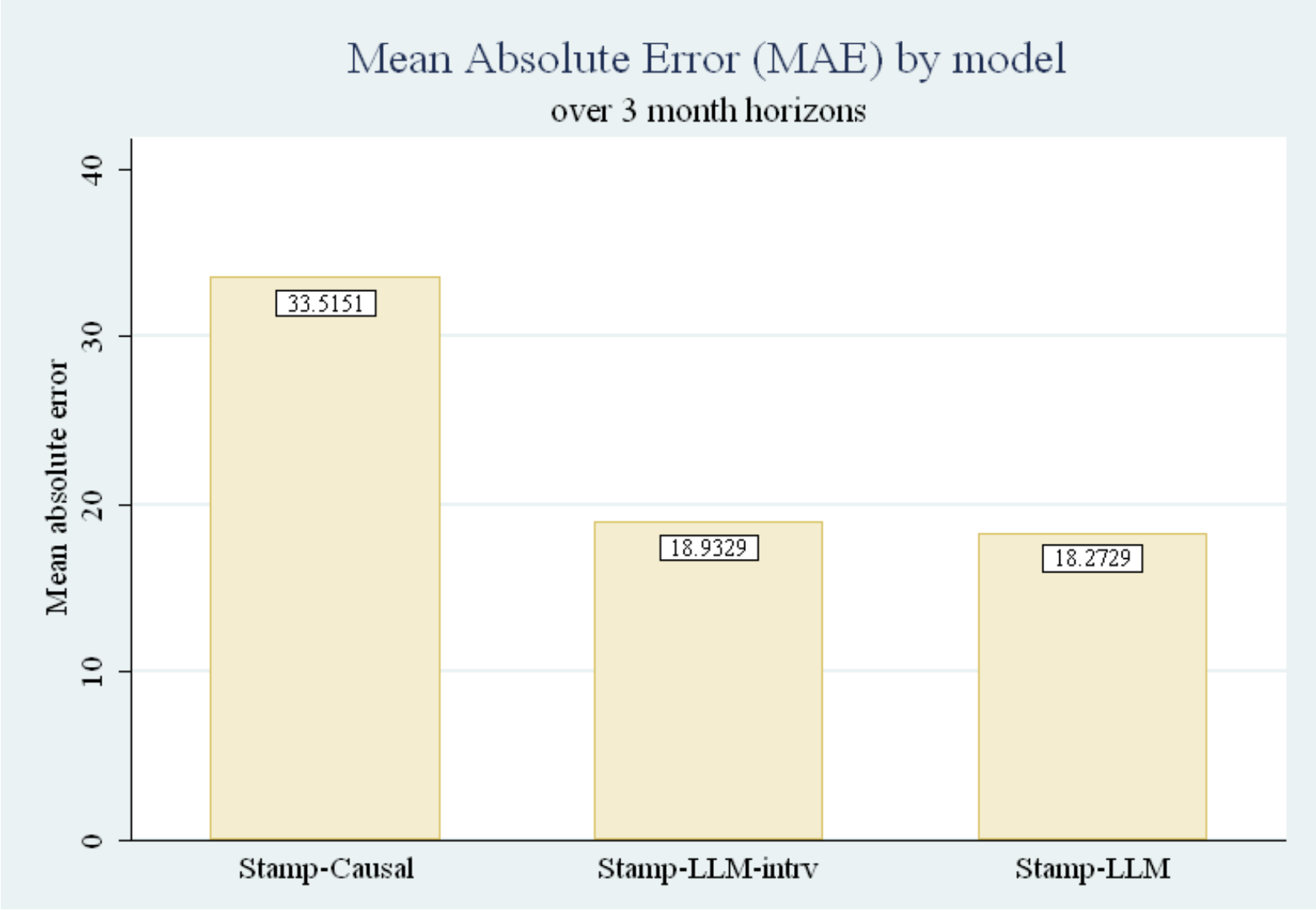
Comparative Forecast Accuracy

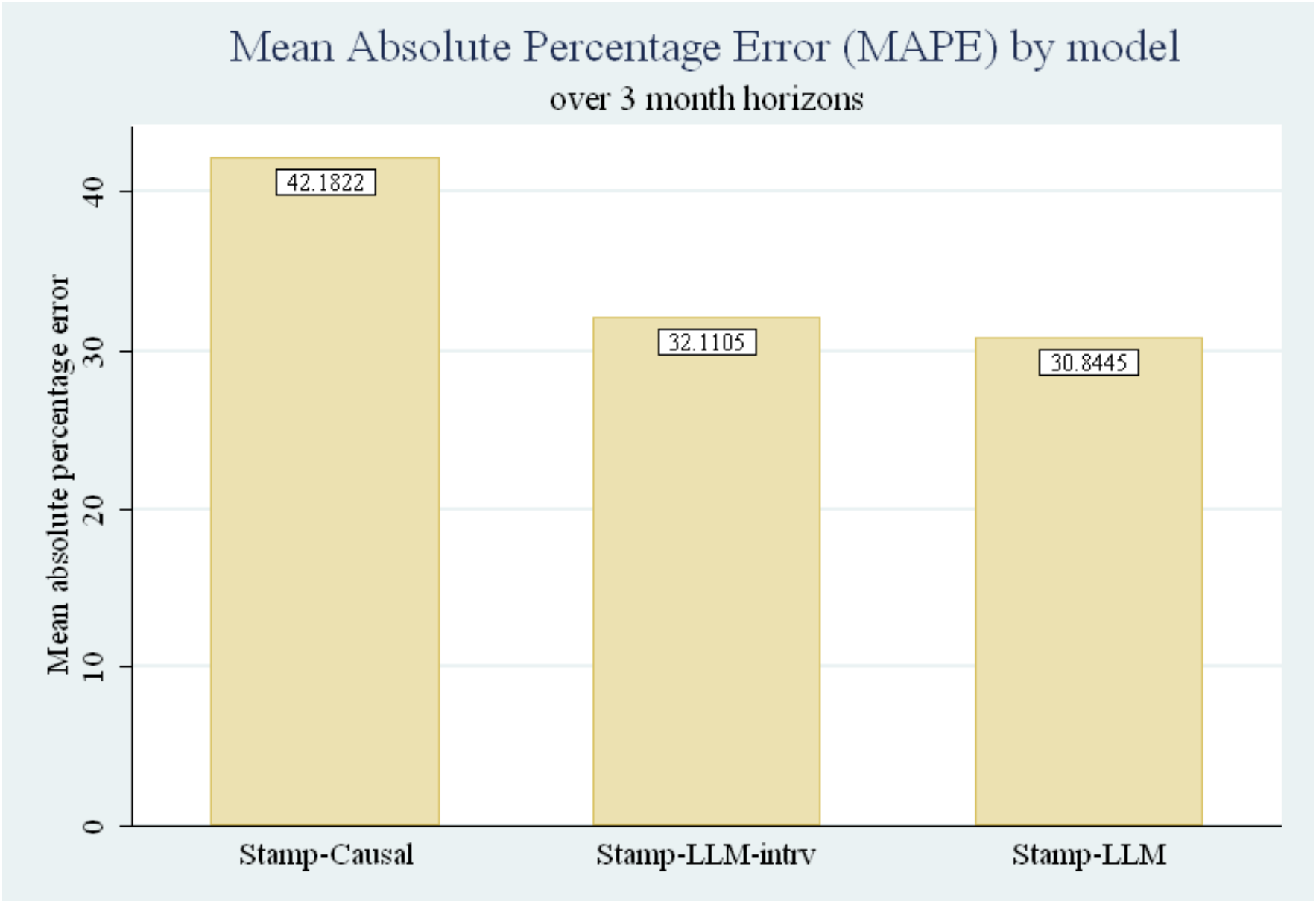
- The univariate models are more accurate than the causal models. This difference does not appear to be significant given the standard errors of the analysis.
- Causal models may induce rigidity in the estimation. So does the density dependence of the series. We now compare 3 state space models with the augmented Kalman Filter.

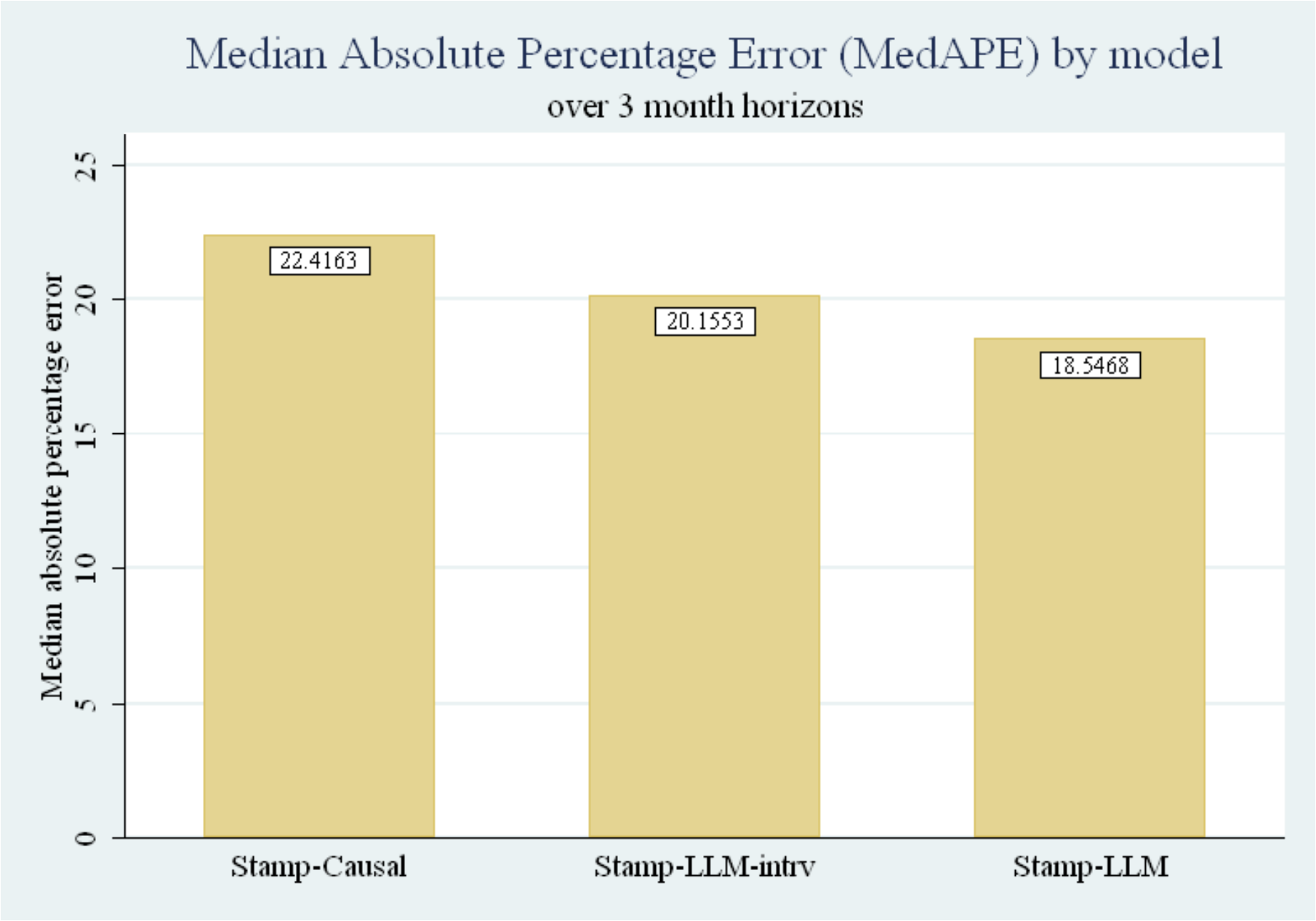
Comparative Forecast Accuracy

over three forecast horizons

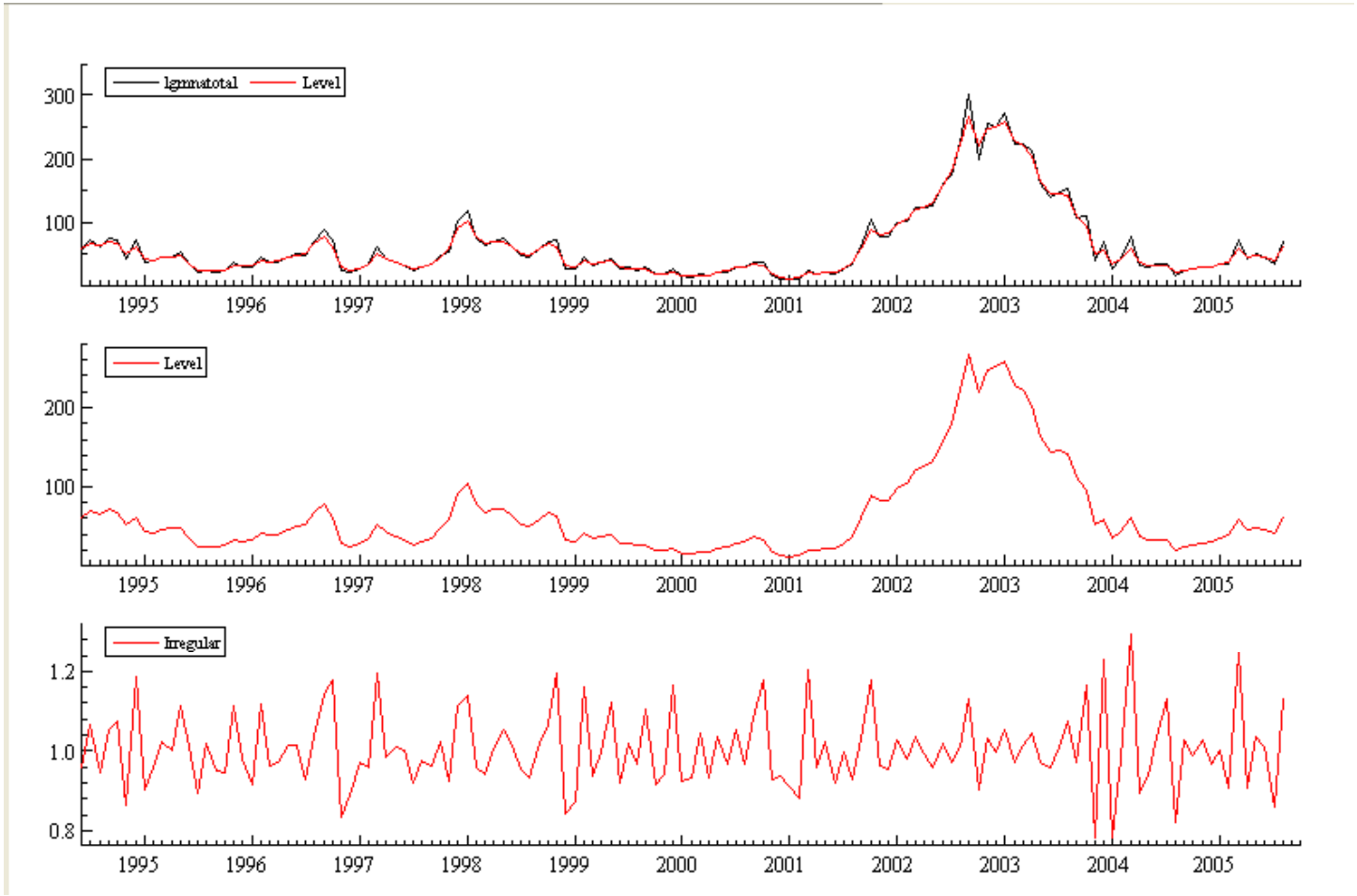
model	MAE	MAPE	MedAPE
Stamp-causal	33.5151	42.1822	22.4163
Stamp-LLM + interv	18.9329	32.1105	20.1553
Stamp LLM	18.2729	30.8445	18.5468


Mean Absolute Error (MAE) by model
over 3 month horizons
Mean absolute error
0
10
20
30
40
33.5151
18.9329
18.2729
Stamp-Causal
Stamp-LLM-intrv
Stamp-LLM

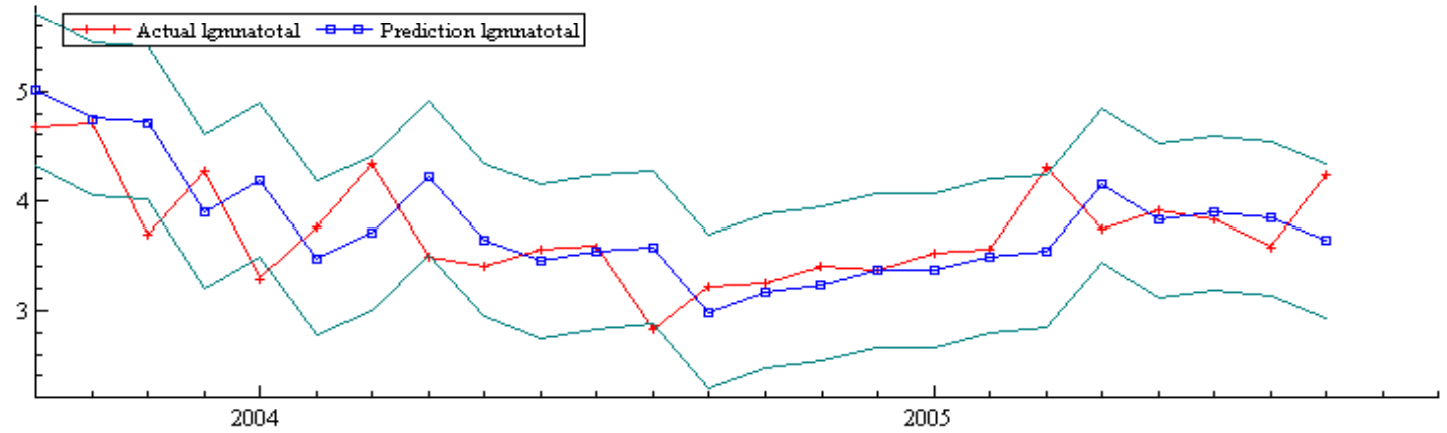

Mean Absolute Percentage Error (MAPE) by model
over 3 month horizons
Mean absolute percentage error
0
10
20
30
40
42.1822
32.1105
30.8445
Stamp-Causal
Stamp-LLM-intrv
Stamp-LLM


Median Absolute Percentage Error (MedAPE) by model
over 3 month horizons
Median absolute percentage error
0
5
10
15
20
25
22.4163
20.1553
18.5468
Stamp-Causal
Stamp-LLM-intrv
Stamp-LLM

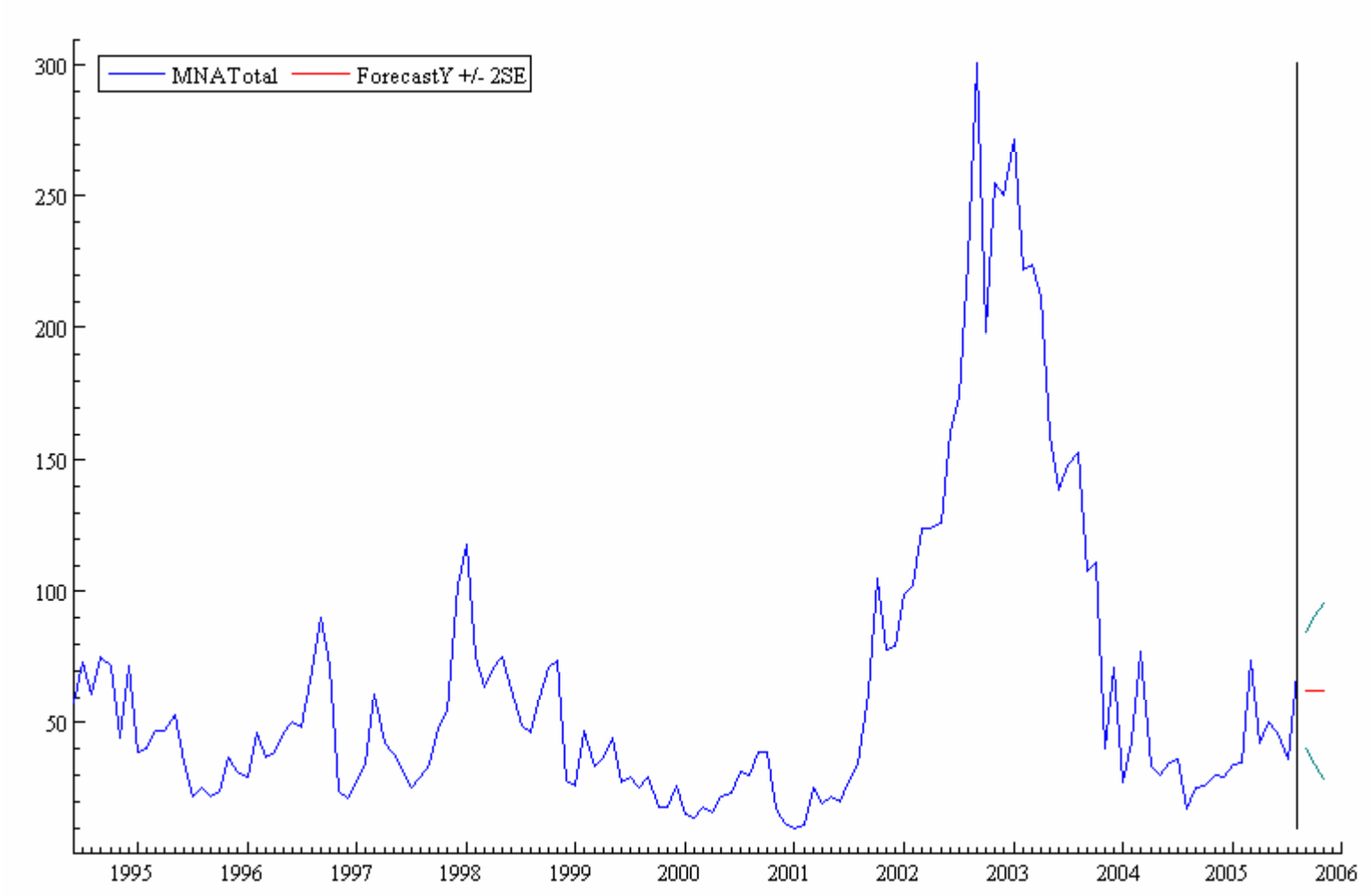
Local Level model Tracking



Higher Resolution of Local Level model Tracking



MNAtotal and Local Level Model



We generate MNAtotal and its forecasts from an anti-log analysis.

Discussion

1. State Space models with readily available weather variables and interventions do not seem to forecast as well as simpler models without those weather variables.
    1. They do not have to forecast the predictors.
    2. Forecasting predictors builds more error into the series.
    3. The final forecast in such a conditional model is based on a flimsier foundation.
2. State space local level models with interventions can predict with more accuracy than the state space causal models.
3. State space local level models appear to be able to forecast MNAtotal very well as the series appears to be characterized by a random walk with noise.

Discussion

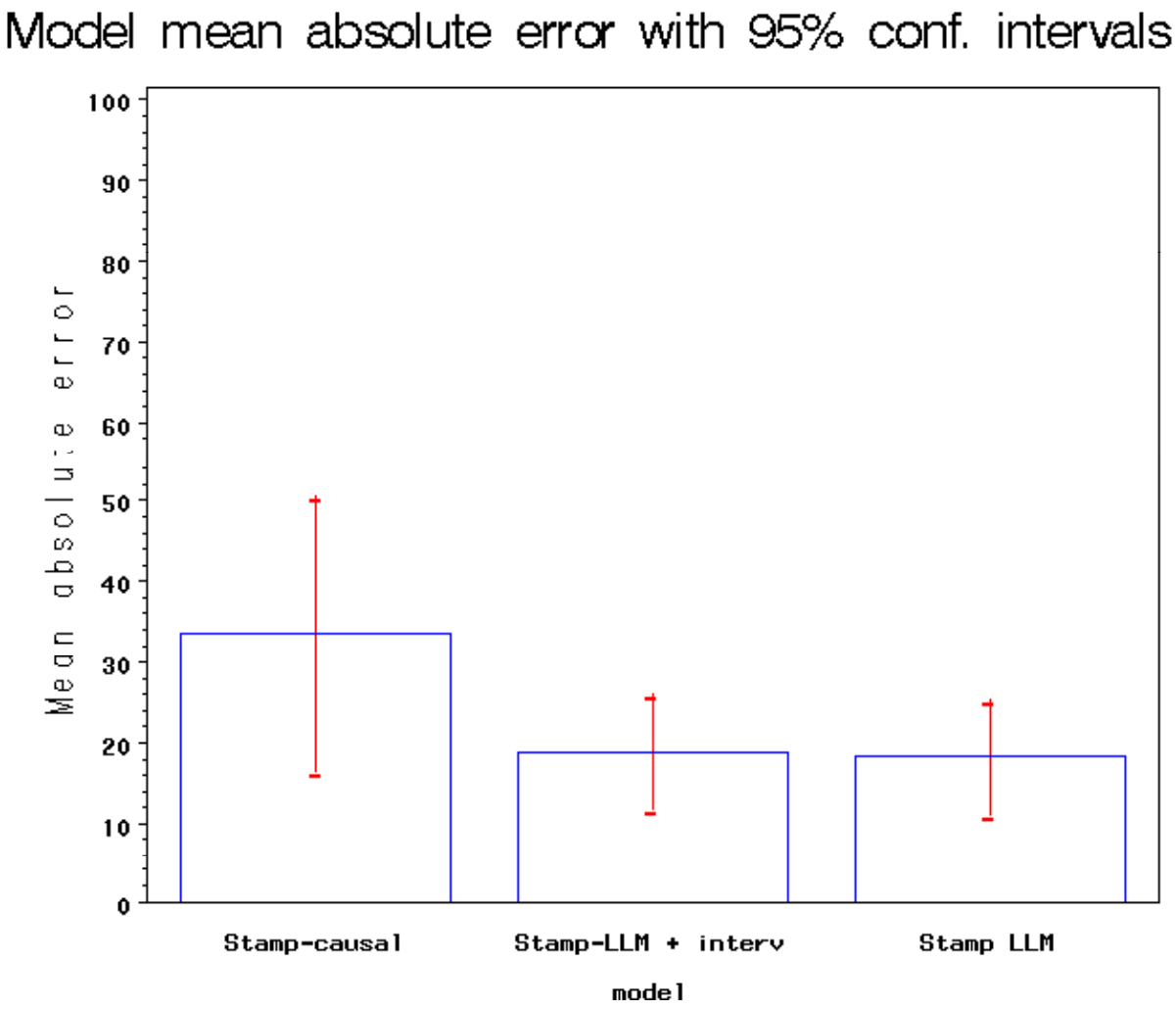
- Of these three state space models, it appears that the local level model exhibits the best forecast performance.
- Interventions are compensated for in the next time period by the Kalman filter correction step.
- The Kalman filter corrects for deviations generated by them in the correction step (from page 29).

$$\alpha_{t+1} = T_t \alpha_t + K_t (y_t - \hat{y}_t)$$

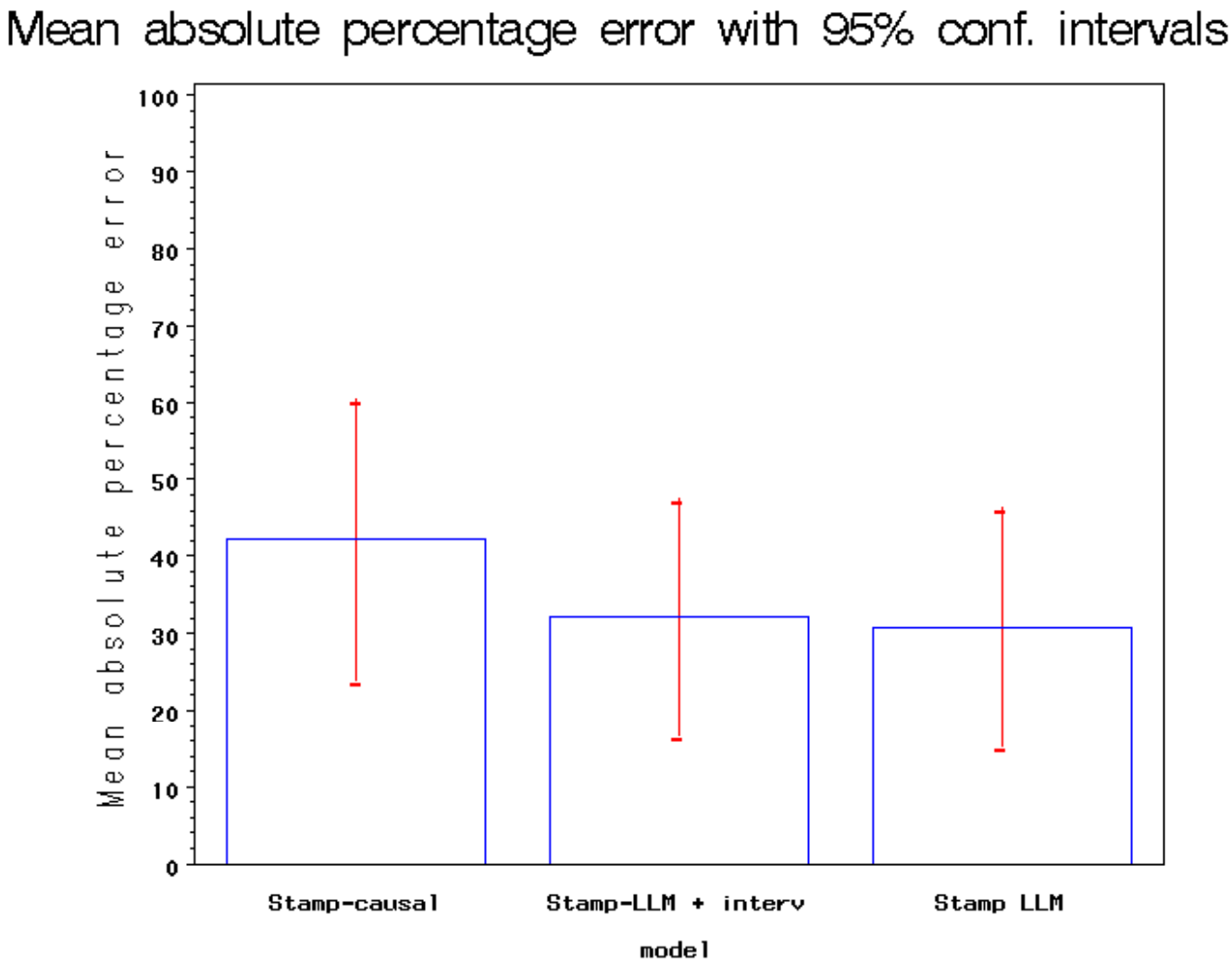
Discussion-cont'd

- Modeling more endogenous lags, although they are significant, rigidifies the autoregressive structure of a local level model. With so many level shifts, the local level needs to maintain flexibility to adjust to the locality of the level (Aoki, M., p. 23).
- Attempts at fixing too much structure on a random walk series can degrade the forecasting accuracy. The single source of error (SSOE) state space model repeatedly specified an MAN (multiplicative errors, additive trend, and no seasonality) model. The series had already been logged. However, there only a partial local trend.
- Simpler models may forecast better. These simpler models include the purely local level model and the local level with interventions model. Modeling interventions does not improve the forecasting significantly. The State Space local level model handles the changes well without the user worrying much about interventions with this series.

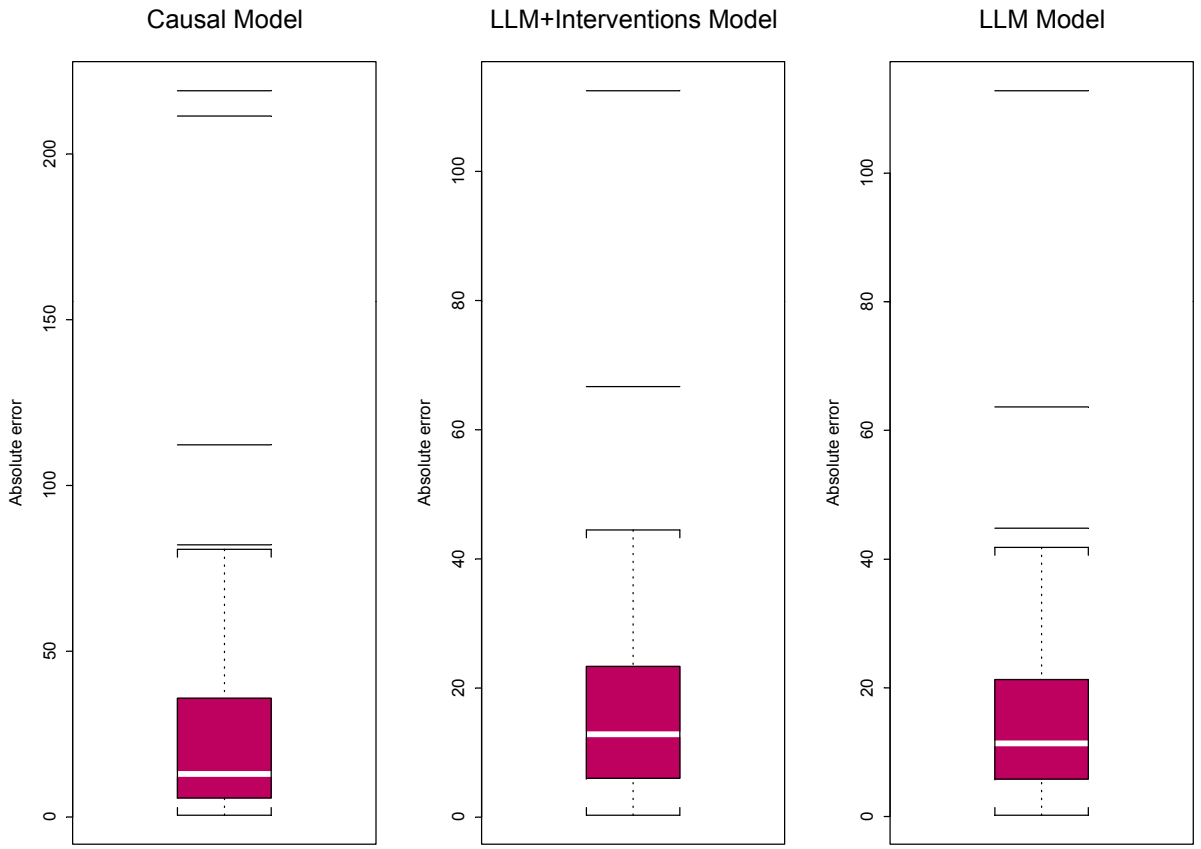
Conclusion: Given the standard error size, we cannot say which model is significantly better.



Conclusion

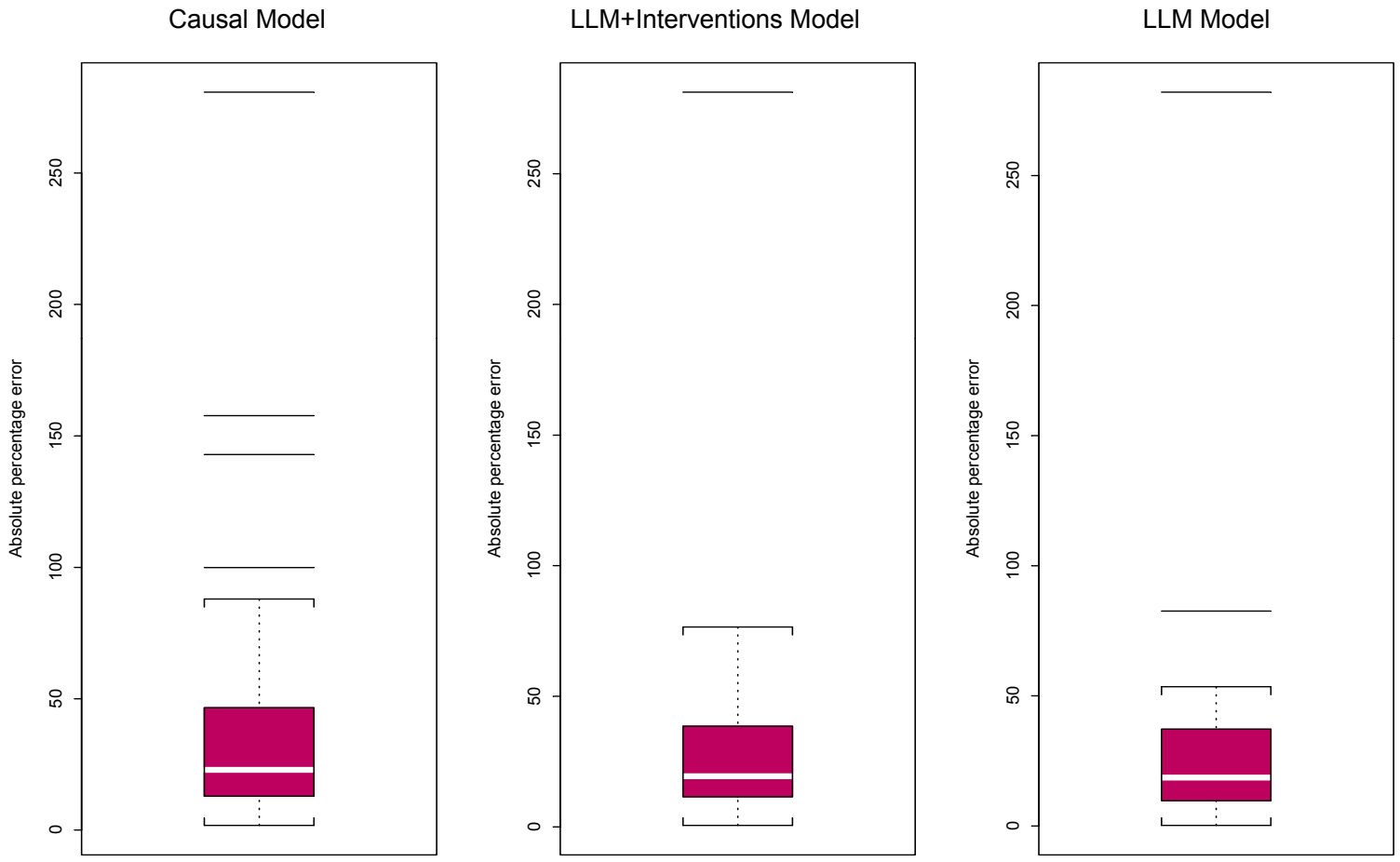

Mean absolute percentage error with 95% conf. intervals
Mean absolute percentage error
100
90
80
70
60
50
40
30
20
10
0
Stamp-causal
Stamp-LLM + interv
Stamp LLM
model

Median Absolute Error



The Median Absolute Error of these causal model appears lower than the others, but the causal model contains more distorting outliers than the others

Median Absolute Percentage Error



The Median Absolute Percentage Error of these models is almost the same, and the causal model contains more distorting outliers than the others

Conclusions 1

1. To answer the first research question, we cannot forecast the abundance of the deer mouse population with readily available weather data well.

   – The MSOE state space model does not seem capable of good forecasting with readily available weather data.
   – Dr. Kenneth Trenbirth from the National Center for Atmospheric research informed me that the weather forecasts are accurate only up to two weeks. Therefore, weather forecasts over a three month forecast horizon are sure to be poor. If forecasts for the predictors over the forecast horizon are sure to be poor, then the forecasts for the "causal" model are sure to be poor.
   – The causal model requires much work and is difficult to forecast because each predictor (and intervention) needs to be forecast before the MNAtotal can be forecast.

Conclusions 2

2. State space models with readily available weather data do not forecast well.
3. State space local level models with interventions can predict with more accuracy than the state space causal models.
4. We can generate a univariate forecast with (MSOE) state space model.

Conclusions 3

5. However, (MSOE) state space local level models appear to be able to forecast MNAtotal very well as the series appears to be characterized by a random walk with noise.

6. The MSOE state space model local level model appears to give the best .

Implications

- When the local level model was compared to the 18 other methods tried to forecast this series, it generated forecast accuracy in *absolute* terms that was superior to all.
- If we were to test these series for significant difference, we might find that *most are not significantly different* from one another in forecast accuracy.

Implications

- Building too much structure into this model impairs the forecasting.
- Modeling the interventions compresses the standard errors and renders the extrapolation likely to overfit
- The local level model is essentially an exponential smoothing process, an extrapolation that combines current data with past estimation.
- The models that forecast best seem to involve some sort of local smoothing built into their algorithm
- The local level model should be used for forecasting. This is the easiest of the models to program.

Limitations

- We cannot commit the universalistic fallacy by saying that what is true for this series is true for all.
- We would have to attempt to test these findings on many different kinds of series to say that these findings can be accepted as general.
- Nevertheless, the findings are heuristic and are suggestive of directions for future research.

What we can conclude

- In any case, the models, whether simple or complex, depend mostly on trapping and recording the minimum number of mice alive.
- To assess this epidemiological risk, given our state of knowledge, trapping and support for it will be necessary.

Appendix 1

- The following Analysis is broken down by cumulative forecast horizon.
- The horizon h is a cumulative forecast horizon.
- The Local level model with dynamic predictors and interventions does not forecast as well as the others.
- The local level model is shown to be somewhat better by all measures, regardless of forecast horizon.

Forecast Accuracy of the models over three forecast horizons (h= number of months)

Mean absolute error

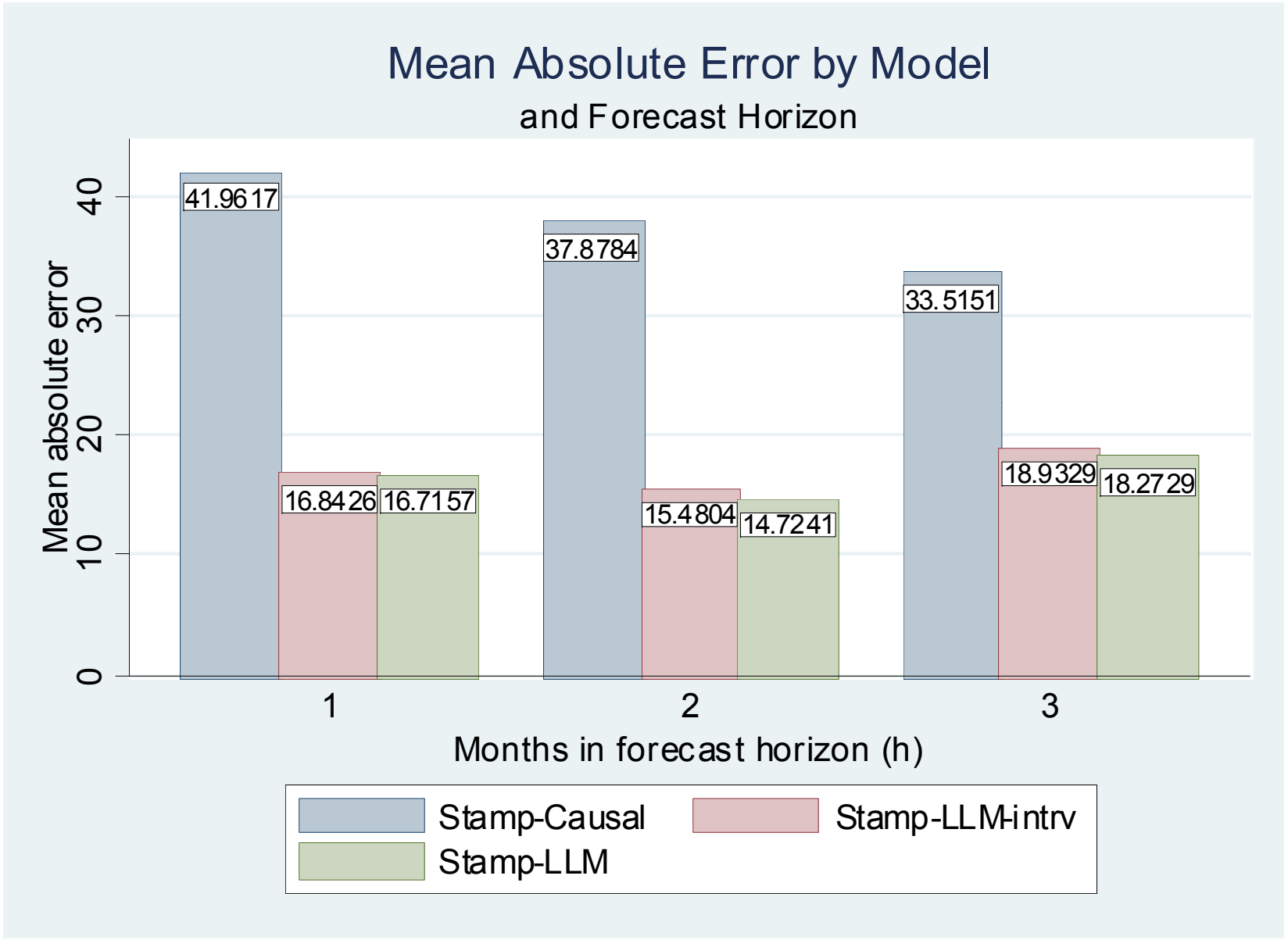
Obs	model	h=1	h=2	h=3
1	Stamp-causal	41.9617	37.8784	33.5151
2	Stamp LLM+interv	16.8426	15.4804	18.9329
3	Stamp-LLM	16.7157	14.7241	18.2729

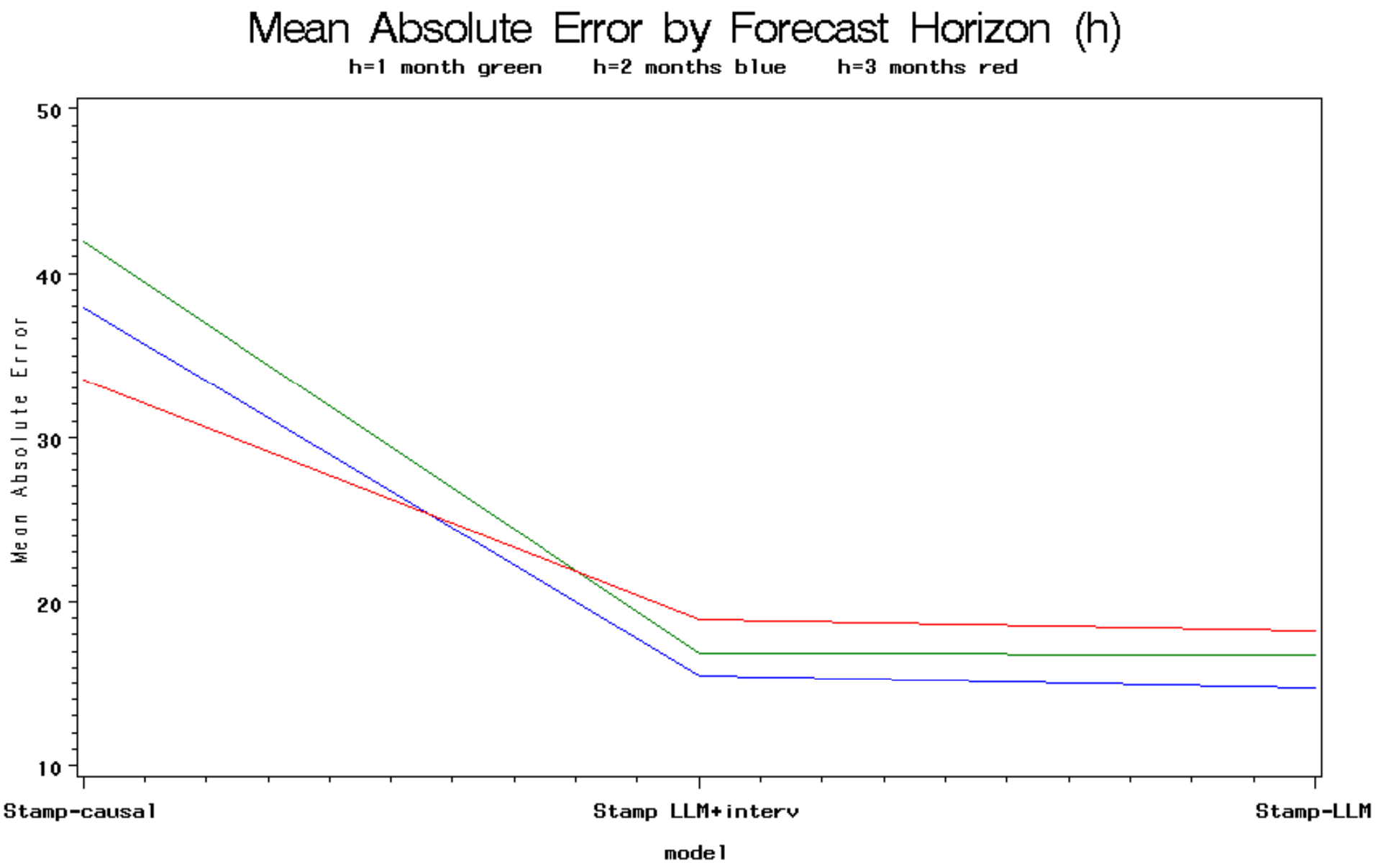
Mean Absolute Percentage Error

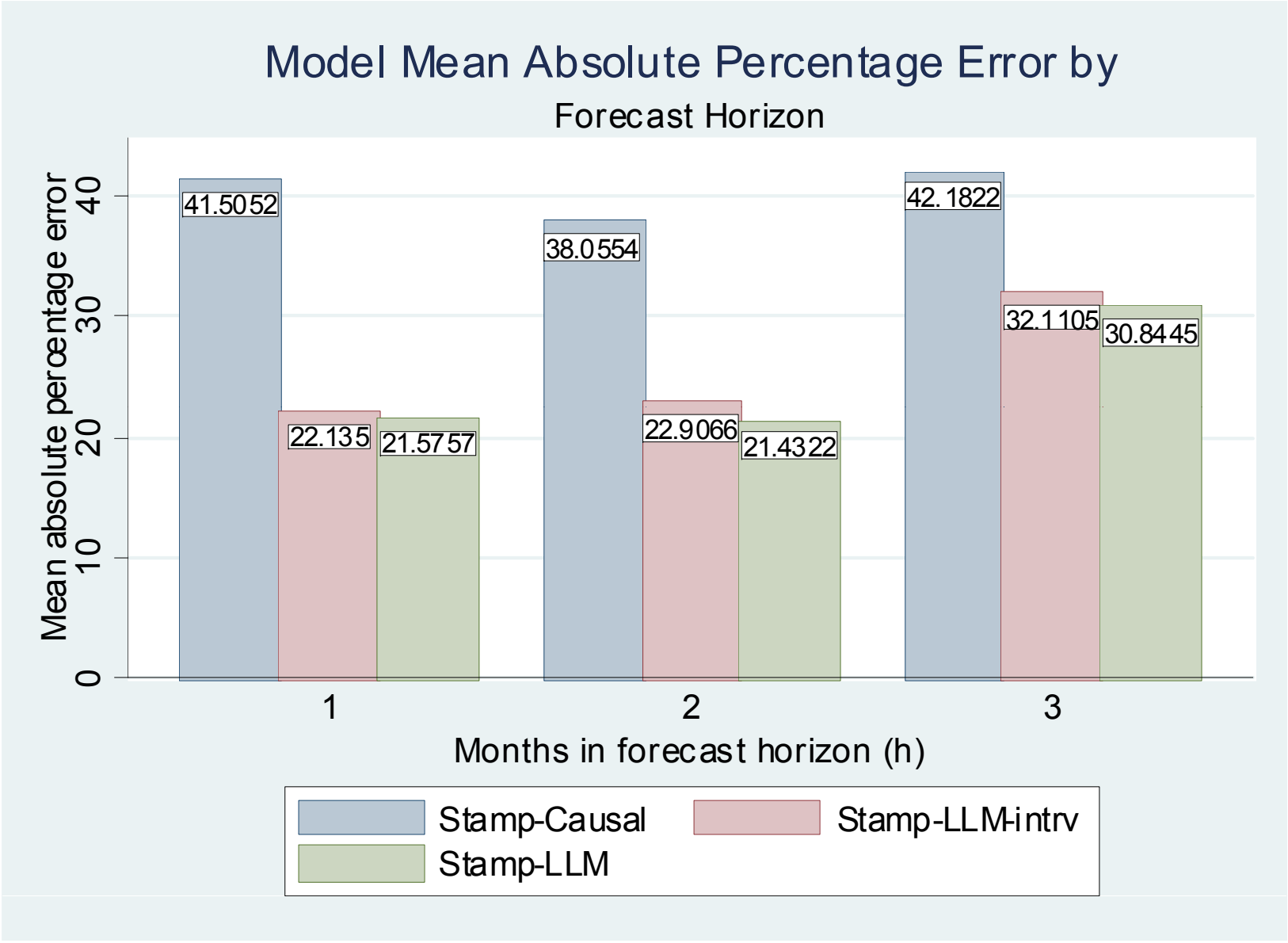
Obs	model	h=1	h=2	h=3
1	Stamp-causal	41.5052	38.0554	42.1822
2	Stamp LLM+interv	22.1350	22.9066	32.1105
3	Stamp-LLM	21.5757	21.4322	30.8445

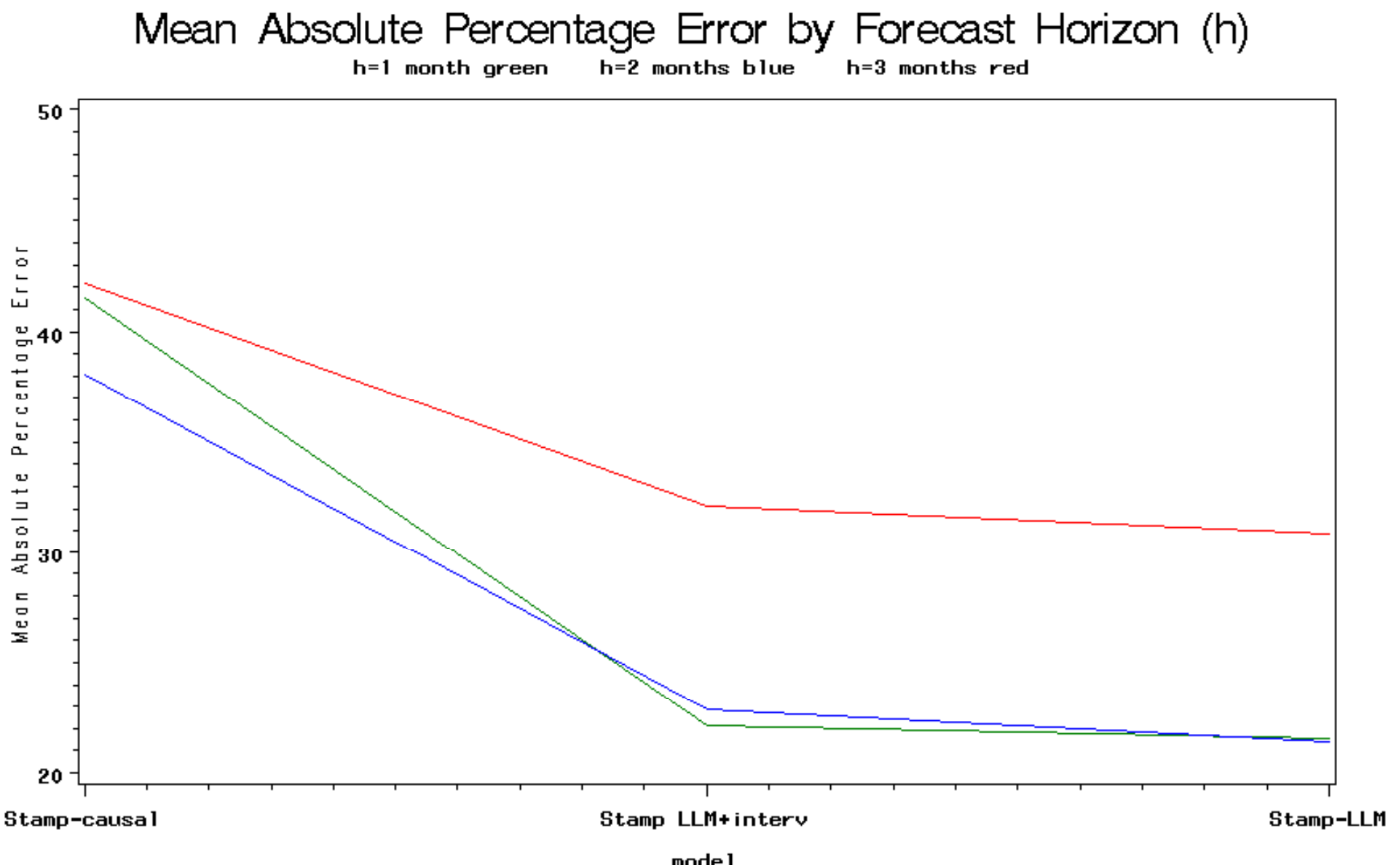
Median Absolute Percentage Error

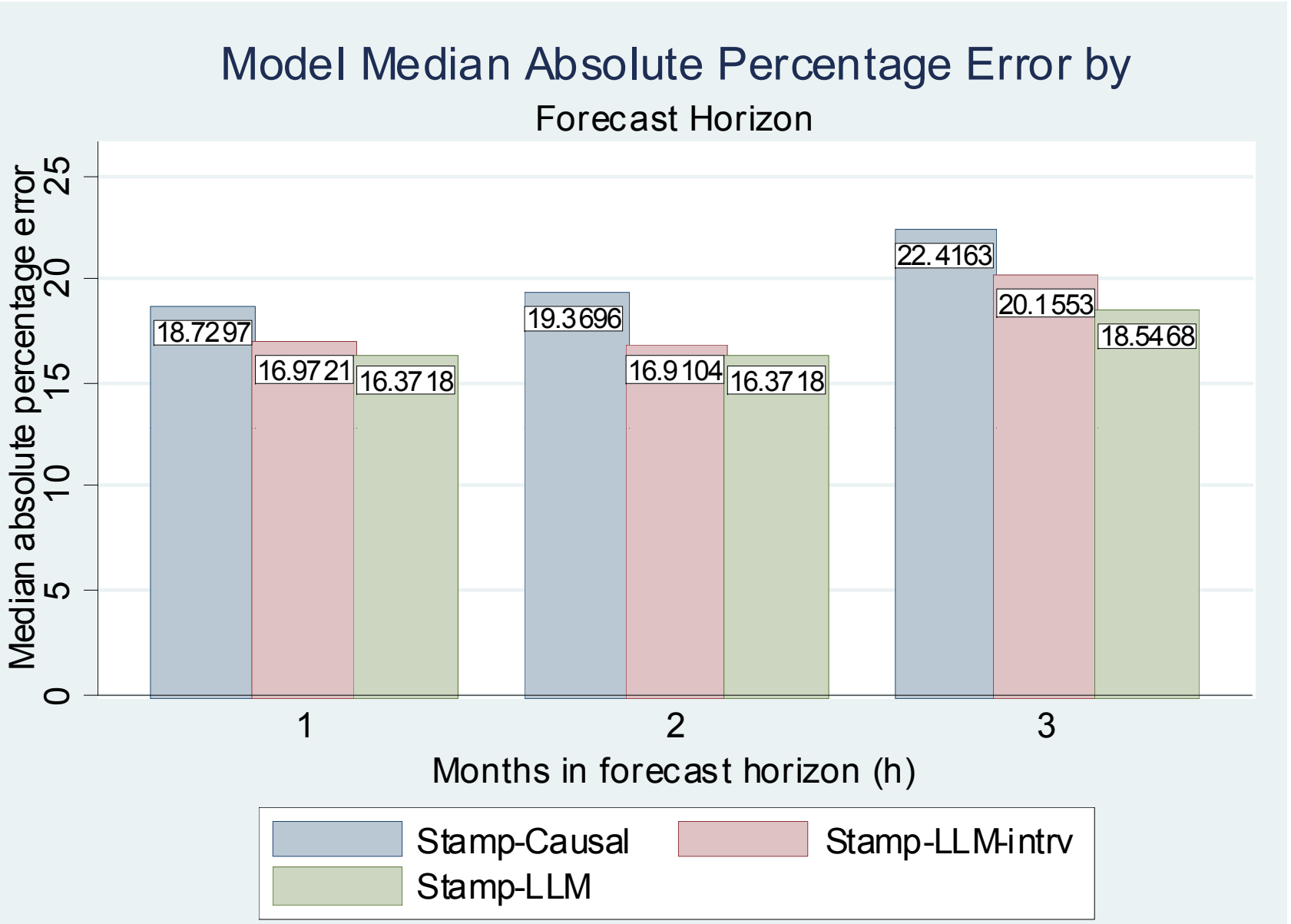
Obs	model	h=1	h=2	h=3
1	Stamp-causal	18.7297	19.3696	22.4163
2	Stamp LLM+interv	16.9721	16.9104	20.1553
3	Stamp-LLM	16.3718	16.3718	18.5468

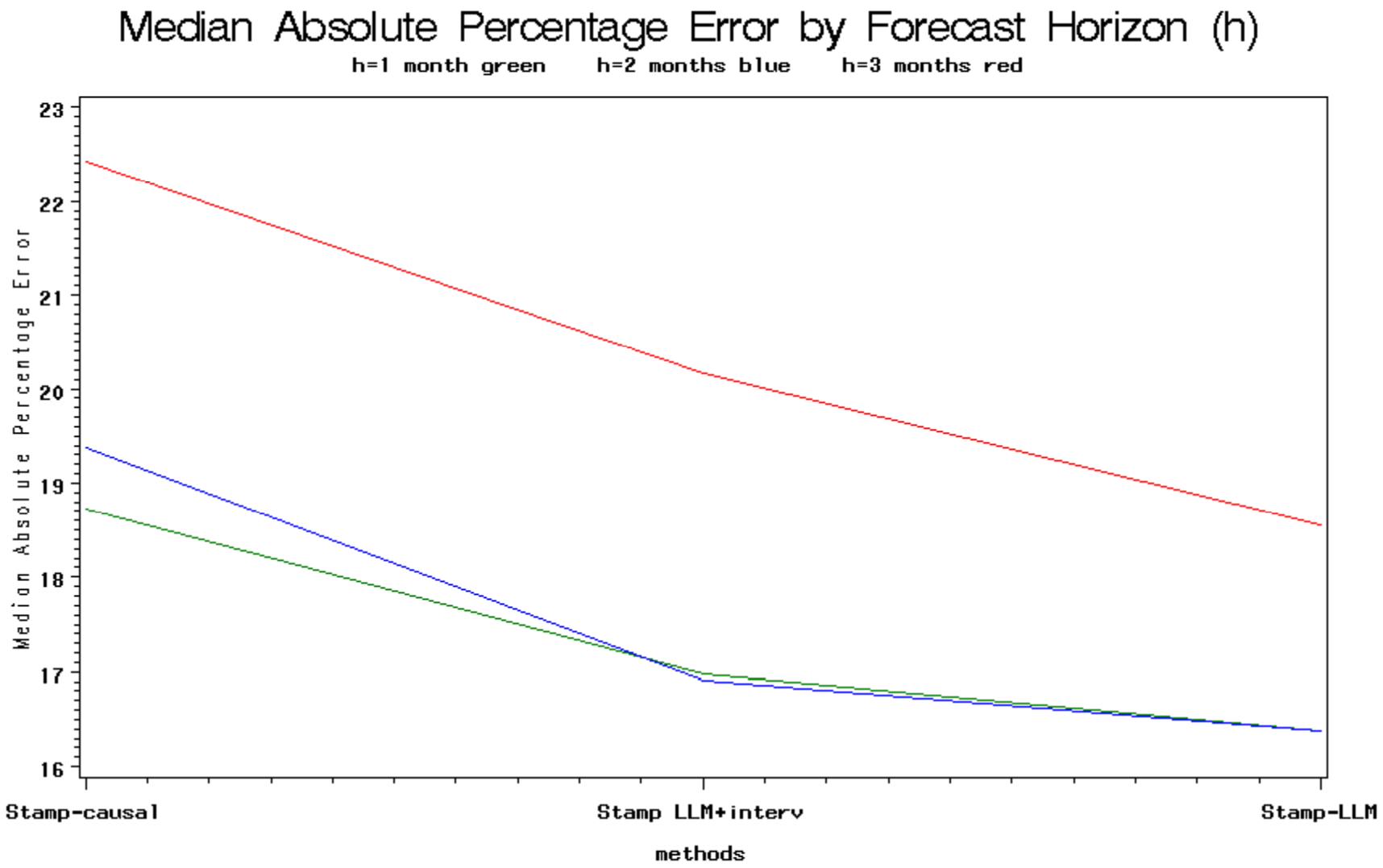

Mean Absolute Error by Model
and Forecast Horizon
Mean absolute error
0
10
20
30
40
41.9617
16.8426
16.7157
37.8784
15.4804
14.7241
33.5151
18.9329
18.2729
1
2
3
Months in forecast horizon (h)
Stamp-Causal
Stamp-LLM-intrv
Stamp-LLM


Mean Absolute Error by Forecast Horizon (h)
h=1 month green
h=2 months blue
h=3 months red
50
40
30
20
10
Mean Absolute Error
Stamp-causal
Stamp LLM+interv
Stamp-LLM
model


Model Mean Absolute Percentage Error by
Forecast Horizon
Mean absolute percentage error
0
10
20
30
40
41.5052
22.135
21.5757
38.0554
22.9066
21.4322
42.1822
32.1105
30.8445
1
2
3
Months in forecast horizon (h)
Stamp-Causal
Stamp-LLM-intrv
Stamp-LLM


Mean Absolute Percentage Error by Forecast Horizon (h)
h=1 month green h=2 months blue h=3 months red
Mean Absolute Percentage Error
50
40
30
20
Stamp-causal
Stamp LLM+interv
Stamp-LLM
model


Model Median Absolute Percentage Error by
Forecast Horizon
Median absolute percentage error
0
5
10
15
20
25
18.7297
16.9721
16.3718
19.3696
16.9104
16.3718
22.4163
20.1553
18.5468
1
2
3
Months in forecast horizon (h)
Stamp-Causal
Stamp-LLM-intrv
Stamp-LLM


Median Absolute Percentage Error by Forecast Horizon (h)
h=1 month green h=2 months blue h=3 months red
Median Absolute Percentage Error
23
22
21
20
19
18
17
16
Stamp-causal
Stamp LLM+interv
Stamp-LLM
methods

References

- Ansley, C.F. and Kohn, R. (1985). Estimation, Filtering and Smoothing in State Space Models with Incompletely specified Initial Conditions. The Annals of Statistics, (13), 1286-1316.
- Aoki, M. (1990) State Space Modeling of Time Series. Springer: Berlin, 23.
- Armstrong, J.S. and Collopy, F (1992) Error Measures for generalizing about forecasting methods: empirical comparisons, IJF (4), 69-90.
- Armstrong, J.S. "Evaluating Forecasting Methods," in Armstrong, J.S. (ed). Principles of Forecasting, Kluwer Academic Publishers: Norwalk, MA. 444-472.
- Auslender, L.E. (1998). Alacart: Poor Man's Classification Trees. SUGI 1998. SAS Institute. Cary: NC.
- Autobox User's Guide v.6. (2006). Automatic Forecasting Systems, Inc., Hatboro. Pa.
- Brieman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees. Chapman and Hall/CRC: New York.
- CART User's Guide (1997). Salford Systems, San Diego, CA.

References-cont'd

- Chitty D, Phipps E, 1966. Seasonal changes in survival in mixed populations of two species of vole. J. Animal Ecol. 35:313-331.
- Clements, M. and Hendry, D.F.(eds). (2004). A Companion to Economic Forecasting. Blackwell Publishing: Malden, MA.
- DeJong, P. (1991). The Diffuse Kalman Filter. The Annals of Statistics (19), 1073-1083.
- DeJong, P. & Penzer, J. (2004). Statistics and Probability Letters (70), 119-125.
- Durbin, J. and Koopman, S.J. (2001) Time Series Analysis by State Space Methods. Oxford University Press: New York.
- Goodwin, P. & Lawton, R (1999). On the asymmetry of symmetric MAPE. IJF (15) 405-408.

References-cont'd

- Hamilton, J.D. (1994). Time Series Analysis. Princeton University Press: Princeton, NJ., 372-408.
- Harvey, A. (1992). Forecasting Structural Time Series Models and the Kalman Filter, Cambridge University Press: New York, Chapter 2.
- Harvey, A. (1993) Time Series Models. MIT Press: Cambridge, MA, Chapters 1 through 4.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001).. The Elements of Statistical Learning. Springer: New York, 266-296.
- Hendry, D. F. and Krolzig, H.-M. 2001. Automatic Econometric Model Selection using PcGets. Version 1.0. User's Guide, Timberlake Consultancy, Ltd: London, UK.
- Koopman, Harvey, Shephard and Ericsson.(2006) Stamp 7. Structural Time Series Analyser, Modeler and Predictor. Timberlake Consultancy, Ltd.: London.
- Krebs, C.J. (2001). Ecology: The Experimental Analysis of Distribution and Abundance. 5th ed. Benjamin Cummings, Menlo Park, California.
- Lutkepohl, H. (1993). Introduction to Multiple Time Series Analysis, 2nd ed. Springer: New York, 415-444.
- Maindonald, J. and Braun J. (2003). Data Analysis and Graphics Using R. Cambridge University Press: New York: NY., 242-252.

References-cont'd

- Mathews, B.P., & Diamontopoulos, A. (1994). Toward a taxonomy of forecast error measures: A factor-comparative investigation of forecast error dimensions, JOF (13), 409-416.
- Meinhold, R.J. and Singpurwalla, N.D. (1983). Understanding the Kalman Filter. The American Statistician,(37). No 2. (May), 123-127.
- Ord, K., Koehler, A. and Snyder, R. (1997) Estimation and Prediction for a Class of Dynamic Nonlinear Statistical models. JASA, Dec. 1997, 1621-1629.
- Proietti, T. (2004). Forecasting with Structural Time Series Models. In Clements, M. and Hendry, D.F. (eds) A Companion to Economic Forecasting. Blackwell Publishing: Oxford,105-133.
- Rosenberg, B. (1973). Random Coefficient Models: the Analysis of Cross-section Time Series by Stochastically Convergent Parameter Regression. Annals of Economic and Social Measurement, 2.,, 399-428, cited in Durbin and Koopman (2000).115-117.
- Snyder, R. and Forbes, C. S. (2002). *Reconstructing the Kalman Filter for Stationary and Non Stationary Time Series.* http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2002/wp14-02.pdf.

References-cont'd

- Stata 9 (2006) Time Series Reference Manual, StataCorp. College Station, Tx.
- Stellwagon, E. and Goodrich's. (2006). Forecast Pro User's Guide, Business Forecast Systems, Belmont, MA.
- Steinberg, D. (2001) .CART: Tree Structured Non-Parametric Data Analysis. Manual. Salford Systems, Inc. San Diego, Ca., 23-27.
- Stroble, C., Zeileis, A. Boulesteix, AL, and Hothorn, T. (2005) Bias in Variable Importance Measures of Ensemble Methods Based on Classification Trees. World Wide Web. http://www.r-project.org/useR-2006/Abstracts/Strobl+Zeileis+Boulesteix+Hothorn.pdf.
- Taylor, J.W. (2003). Exponential Smoothing with a Damped Multiplicative Trend. International Journal of Forecasting (19).715-725.
- Welch, G. and Bishop, G. (2006). An Introduction to the Kalman Filter. University of North Carolina Computer Science Dept.: Chapel Hill: World Wide Web: http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf.
- West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models, 2nd ed., Springer: New York.

References-cont'd

- Yaffee, R. A. (in prep). Introduction to Forecasting Time Series using Stata. StataCorp. College Station, TX., Chapter 2.
- Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. Journal of the American Statistical Assn. Vol. 93. No. 441. (Mar. 1998), 120-131.
- Zivot, E. and Wang, J. (2006). Modeling Financial Time Series with S-Plus. 2nd ed. Springer: New York., 519-569.
- Zivot, E., Wang, J. and Koopman, SJ. (2004). State Space Modeling in Macroeconomics and Finance using SsfPack in S+Finmetrics. In Harvey, A., Koopman, SJ, and Shepard, N. (eds). State Space and Unobserved Components Models: Theory and Applications, Cambridge University Press: New York, 287.

E-mail addresses

- Robert A. Yaffee, Ph.D. yaffee@nyu.edu
- Kent D. Wagoner, Ph.D. kwagoner@ithaca.edu
- Rick J. Douglass, Ph.D. rdouglass@mtech.edu
- Brian R. Amman, Ph.D. cxx1@cdc.gov
- Tom Ksiazek, Ph.D. tksiazek@cdc.gov
- James N. Mills, Ph.D. jum0@cdc.gov
- Kostas Nikolopoulos, Ph.D., kostas.nikolopoulos@mbs.ac.uk
- David Reilly, dave@autobox.com
- Sven F. Crone s.crone@lancaster.ac.uk