Statistical modeling for prospective surveillance: paradigm, approach, and methods

Al Ozonoff*, Paola Sebastiani

Boston University School of Public Health
Department of Biostatistics

aozonoff@bu.edu

3/20/06

Prospective surveillance

Surveillance defined

"Surveillance is the cornerstone of public health practice." (Thacker, 2004)

- **Surveillance**: "The systematic collection, consolidation, analysis and dissemination of data in public health practice." (Langmuir, 1963)
- "The ongoing systematic collection, analysis, and interpretation of outcome-specific data for use in the planning, implementation, and evaluation of public health practice." (Thacker, 2000)
- Broad definition supports a wide variety of surveillance practices.

Traditional surveillance

Traditional practice of surveillance has nearly 400 years of history.

- Focus on retrospective examination of data.
- Infectious disease: basis for outbreak investigation.
- Other health outcomes: allows study of trends and evaluation of policy changes; control measures; public health practice.
- Hypothesis-generating activity.

Prospective surveillance

New (and evolving) paradigm for public health surveillance.

- More timely collection of data.
- Wider range of "outcome-specific data".
- Hypothesis-testing activity.
- Prototype: "syndromic surveillance".
- Principles embodied in newly-formed International Society for Disease Surveillance (ISDS).

Prospective surveillance

Some challenges currently facing prospective surveillance:

- Informatics: speedy (electronic) acquisition of data.
- Anomaly detection: near-real-time identification of outbreaks.
- False alarms: potential hypothesis testing on daily basis requires strict control of Type I error.
- Forecasting: modeling of underlying process for projection of future patterns of disease.

Forecasting

Begin with some model that will yield one-step-ahead prediction.

- Accuracy of forecast will depend on model chosen.
- Fundamental paradigm: first, establish what is "normal". Then, be vigilant for deviations from normal behavior. Focus on behavior of one-step-ahead (or many-step-ahead) residuals.
- For prospective surveillance, measure of forecasting capability is predictive accuracy (e.g. RMSE).

Forecasting

Anomaly detection:

- Relies on one-step-ahead residuals.
- Small residual $\Rightarrow$ "normal" behavior.
- Large residual $\Rightarrow$ deviation from normalcy.
- Performance of baseline model (reduction of residuals) is paramount.
- Relentless pursuit of forecasting ability may lead to models that obscure underlying processes.
- Are such models robust to changing conditions?

Forecasting

Aside from anomaly detection: consider study of disease process, epidemiology/transmission of disease, and long-range forecasting.

- Careful selection of model should yield representation of some aspects of disease process.
- Residuals consist of effects *not explained by model*.
- "Random variability" simply an admission that model does not account for all observed variation.
- Must reach a balance between parsimony/interpretability and performance. Not a new idea!

Forecasting

Problem: life is complicated.

- Bench sciences: make clever choice of experimental design or measurement device.
- Surveillance: constrained by limitations of data. Must be *even more clever*.
- Influenza demonstrates rich, complex dynamics.
- Further confounded by human behavior, environmental factors.

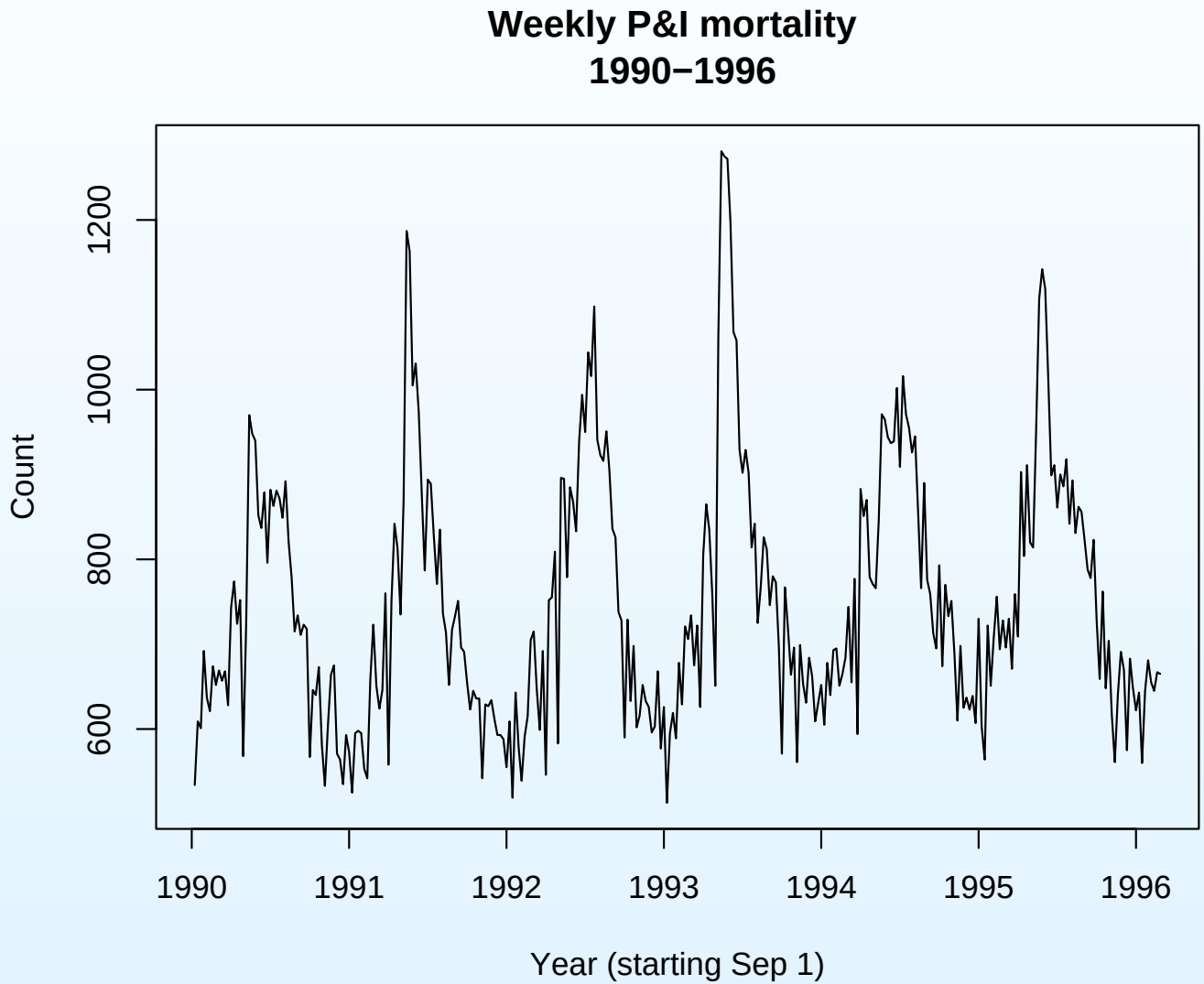
Influenza surveillance

Models for influenza

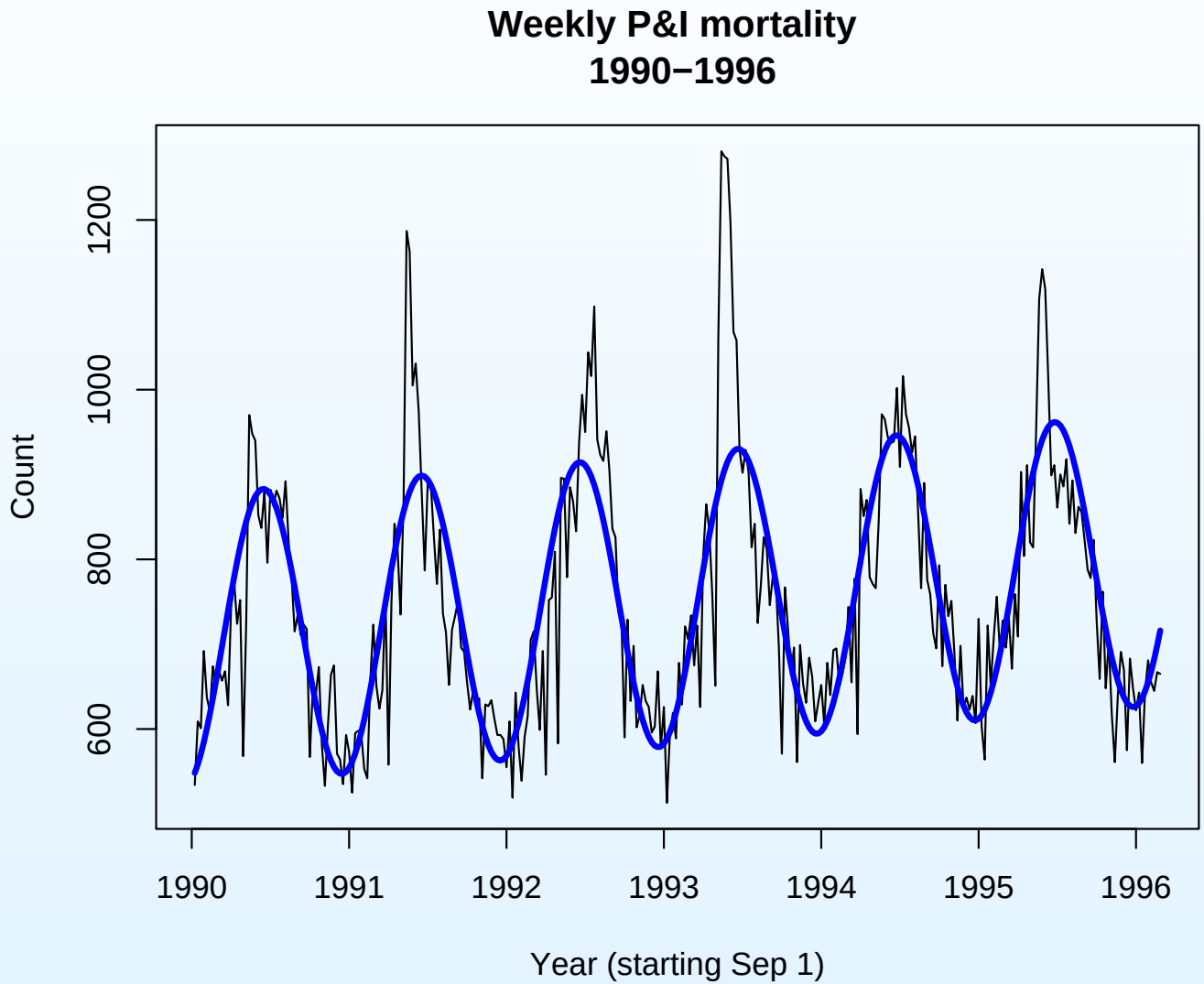
Serfling's method for influenza.

- Traditional approach: model respiratory illness as sinusoid (Serfling's method).
- Problem: sinusoid fits data poorly during epidemic periods (i.e. winter-time increase in flu activity).
- Implication for prospective surveillance: decreased performance (i.e. lower power for detection of outbreaks) during winter months.

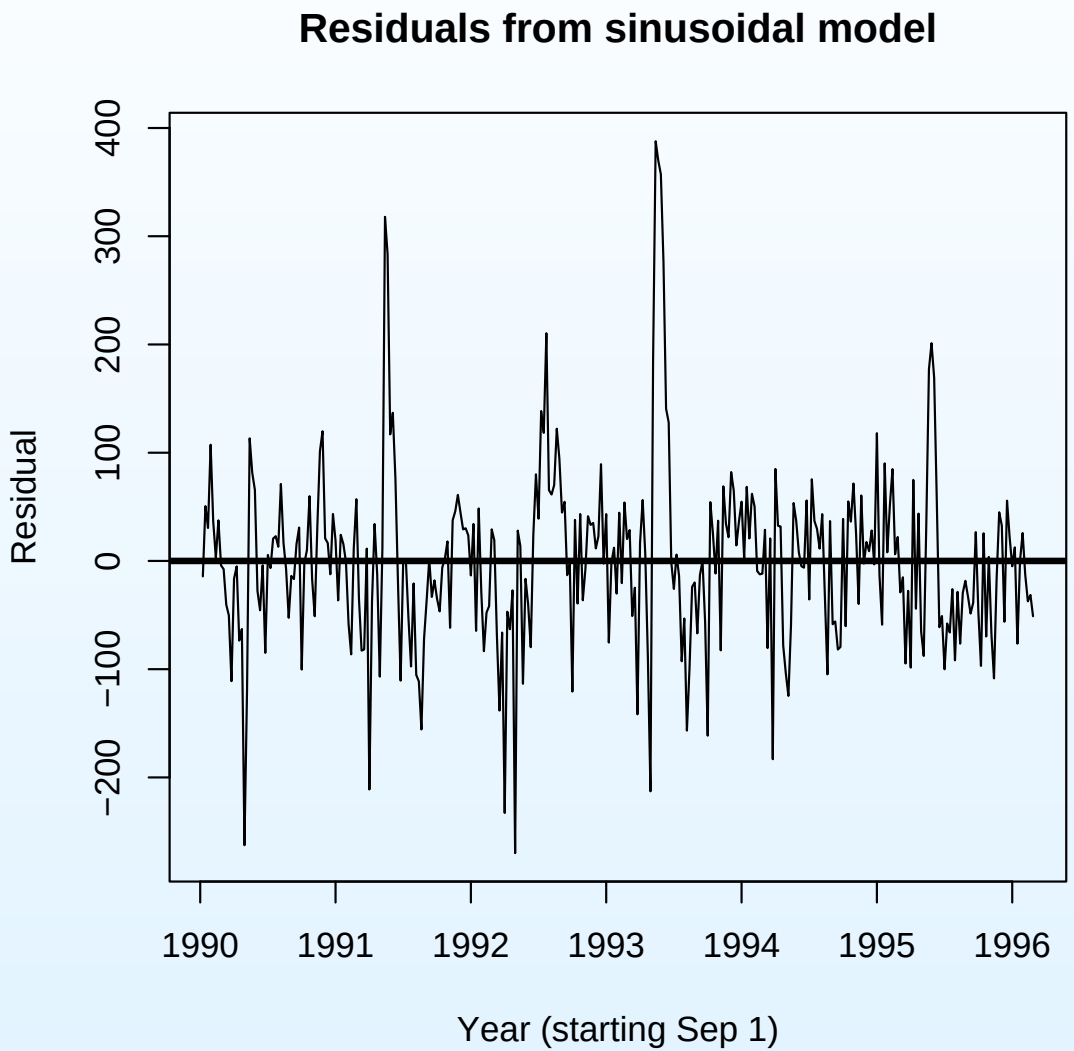
National P+I mortality



National P+I mortality



National P+I mortality



Further motivation

Other reasons for developing more sophisticated models for influenza surveillance data:

- Prospects of novel strain (e.g. “avian flu” H5N1) emerging to cause pandemic illness. Could see new dynamics of transmission, epidemiology.
- Preparedness: understand past pandemics to learn lessons for future events. Focus shifts back to disease process.
- Challenging problem: model spread of disease across space and time. Current univariate models don’t seem to generalize well to spatio-temporal models.
- Seasonality of influenza not completely understood.
- Data sources beyond traditional influenza surveillance data are increasingly becoming available.

Hidden Markov Models

Classical approach

- Serfling's model: underlying seasonal baseline is roughly sinusoidal. May be driven by temp; annual patterns (e.g. school year); dynamics of disease.

$$Y_t = \alpha_0 + \alpha_1 t + \beta_1 \sin\left(\frac{2\pi t}{52}\right) + \beta_2 \cos\left(\frac{2\pi t}{52}\right) + \epsilon_t$$

- Large deviations above this baseline indicate epidemic state. Integrating residuals allows calculation of "excess mortality" i.e. mortality attributed to influenza above what would be expected, accounting for seasonal variation.
- Performs well for what it is asked to do. Not good at one-step-ahead predictions, since model fit is poor during epidemic state.

Other approaches

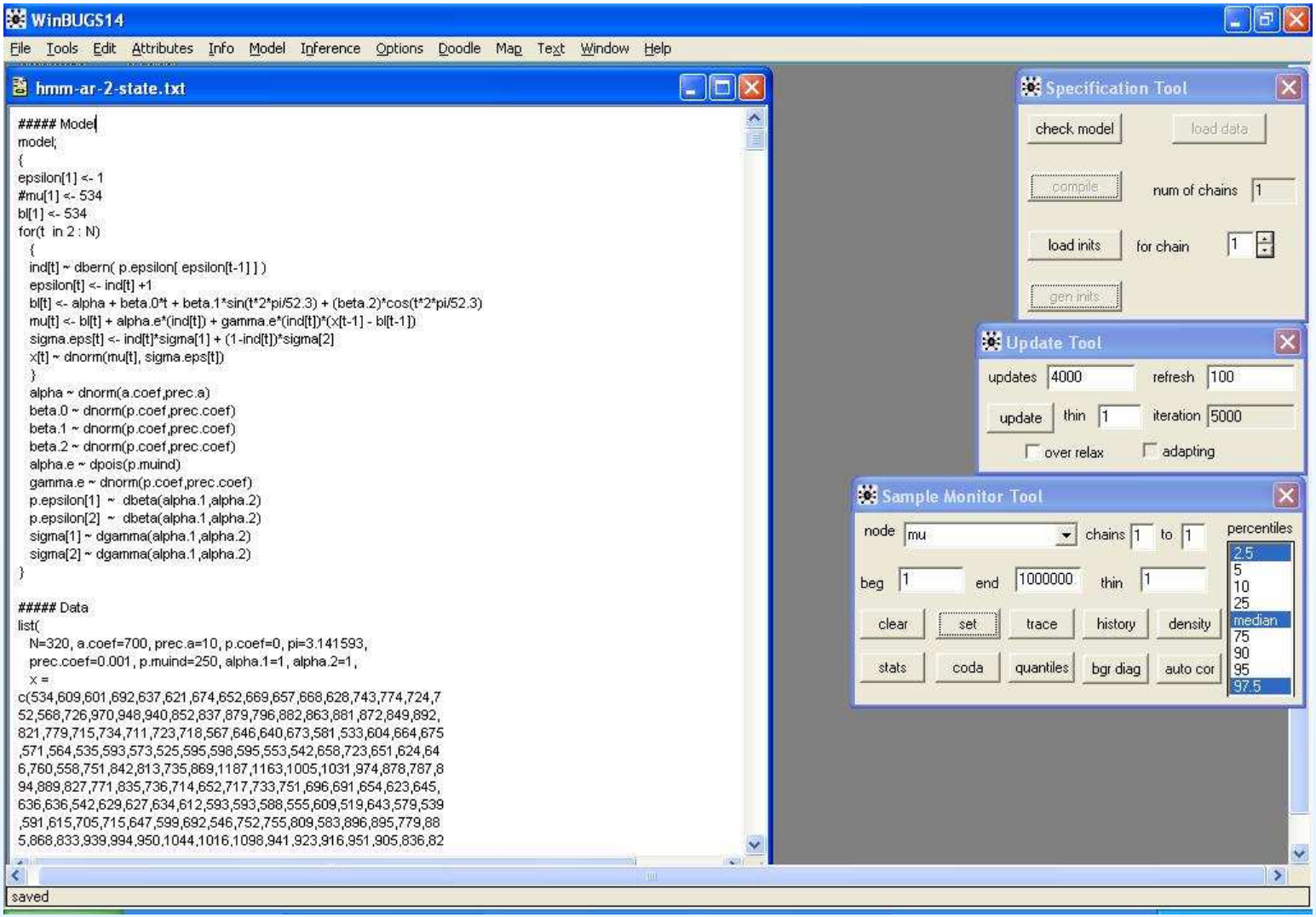
- Periodic regression with auto-regressive component (PARMA). Used in syndromic surveillance settings. Better model fit thanks to AR component.
- "Method of analogues": non-parametric forecasting. Outperforms many methods in one-step-ahead prediction (Viboud 2003). Non-parametric $\Rightarrow$ ignores and obscures knowledge about mechanism of disease.
- Nuño and Pagano developing mixed models approach using Gaussian with phase shift as random effect. Also incorporate bimodal Gaussian for occasional dual-wave behavior.

Hidden Markov Models (HMMs)

Our approach: HMMs.

- 'Hidden" (latent, unobserved) discrete random variable, representing some aspect of disease process.
- Observed variables are modeled, conditional upon the hidden state. Know which state $\Rightarrow$ know distribution of observed random variable.
- Markov property: conditional probability of state change (transition probability) depends only on the value of latent state at previous time point. Thus specify the Markov model for k states with a $k \times k$ matrix of transition probabilities, and the distributions of the observed data conditional on the hidden state.
- Parameter estimation using Bayesian inference Using Gibbs Sampling (BUGS). Freeware available, e.g. WinBUGS, BRUGS.

WinBUGS screen shot

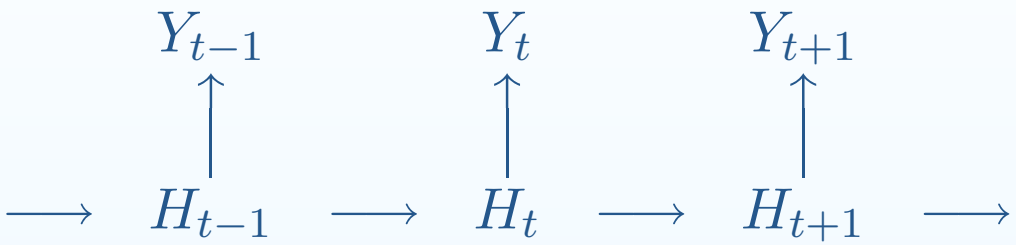


Hidden Markov Models (HMMs)

Model fitting in WinBUGS:

- Sequence of hidden states is treated as a free parameter and fit simultaneously with other model coefficients.
- Computational demanding for long time series; parameter space has order k^n.
- Convergence via Gibbs sampling may be an issue, esp for misspecified models.
- Latent variable provides information about mechanism of disease. Epidemic and non-epidemic behavior can be modeled separately.

Hidden Markov Models (HMMs)



Y_t are observed data i.e. weekly P&I counts.

H_t are the hidden states (for us, 2-state model).

Arrows indicate conditional dependencies.

$$Y_t \sim \alpha_0 + \alpha_1 t + \beta_1 \sin\left(\frac{2\pi t}{52}\right) + \beta_2 \cos\left(\frac{2\pi t}{52}\right) \Big| H_t = 0$$

$$Y_t \sim \left(\alpha_0 + \alpha_{\mathbf{e}}\right) + \alpha_1 t + \beta_1 \sin\left(\frac{2\pi t}{52}\right) + \beta_2 \cos\left(\frac{2\pi t}{52}\right) \Big| H_t = 1$$

Evaluation

Evaluation scheme for outbreak detection:

- Systematically investigate various HMMs and evaluate (with other approaches) using RMSE on one-step-ahead predictions.
- Use fixed period (e.g. 1990-1994) to fit all models, and subsequent year (1995) for predictions. Repeat on other time periods so evaluation is not dependent on time period chosen. “Virtual prospective surveillance” (Seigrist).
- Compare several HMMs; Serfling’s method; PARMA; working on implementation of other methods.

Preliminary results

- Research supported by pilot funds from the Blood Center of Wisconsin. Fourth month of a 10 month funding period; results are preliminary.
- Presenting goodness-of-fit evaluation only; prospective evaluation in progress.
- First step: evaluate HMMs on 122 Cities data.
- Eventually, follow similar approach with influenza-like illness (ILI) data. Goal: predictive spatio-temporal models of influenza morbidity.

Data

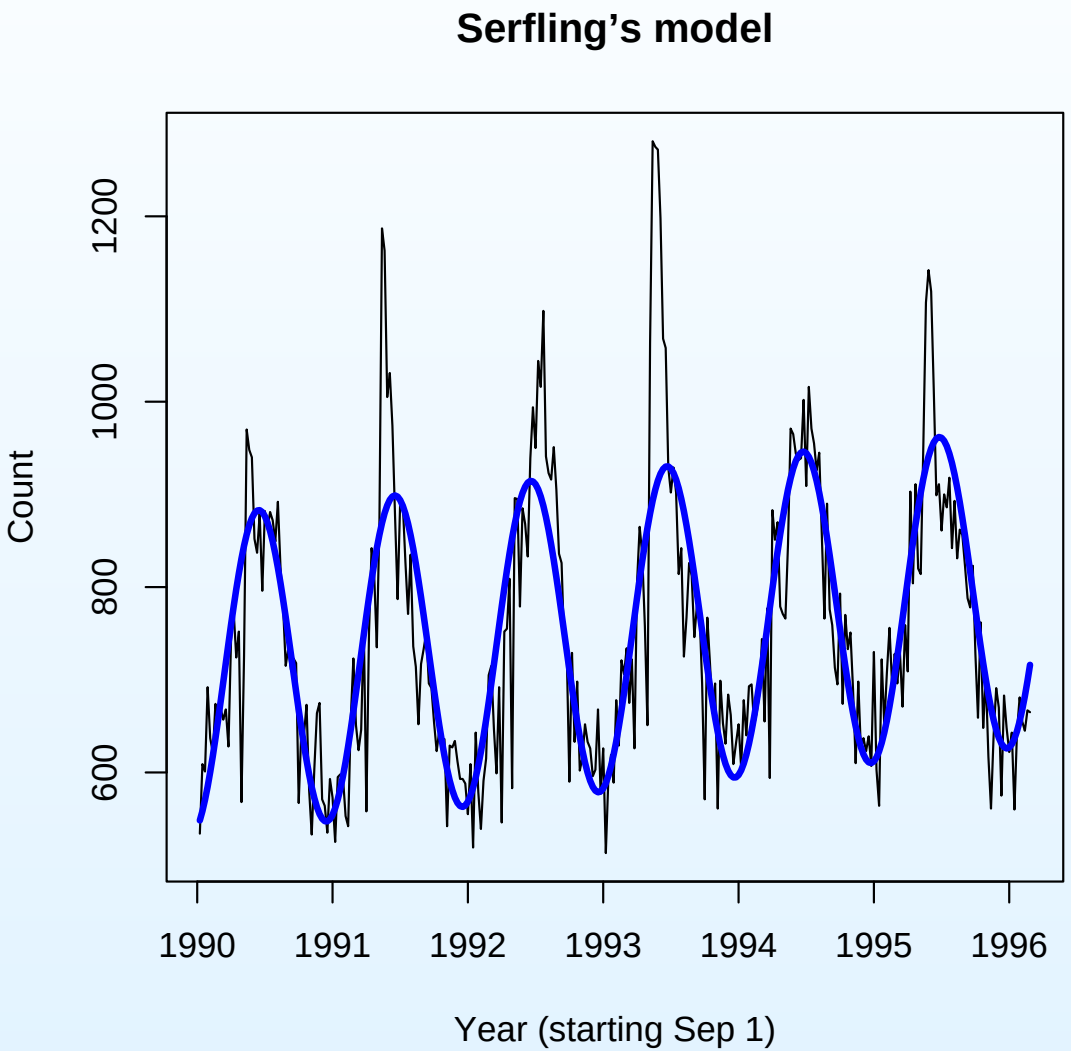
122 Cities influenza surveillance system:

- CDC operated program running continuously since 1962.
- Weekly counts attributed to pneumonia and influenza (P&I). Reporting lag of 2-3 weeks.
- Approx 25% coverage of U.S. pop'n. Used by CDC for determining epidemic influenza (Serfling).
- Age-specific counts available. 122 cities divided into 9 administrative regions, roughly 14 cities per region.
- Limitations: difficult to accurately attribute deaths to influenza; mortality known to lag morbidity (e.g. ILI activity); dynamics may differ from morbidity (depending on circulating viral strains).

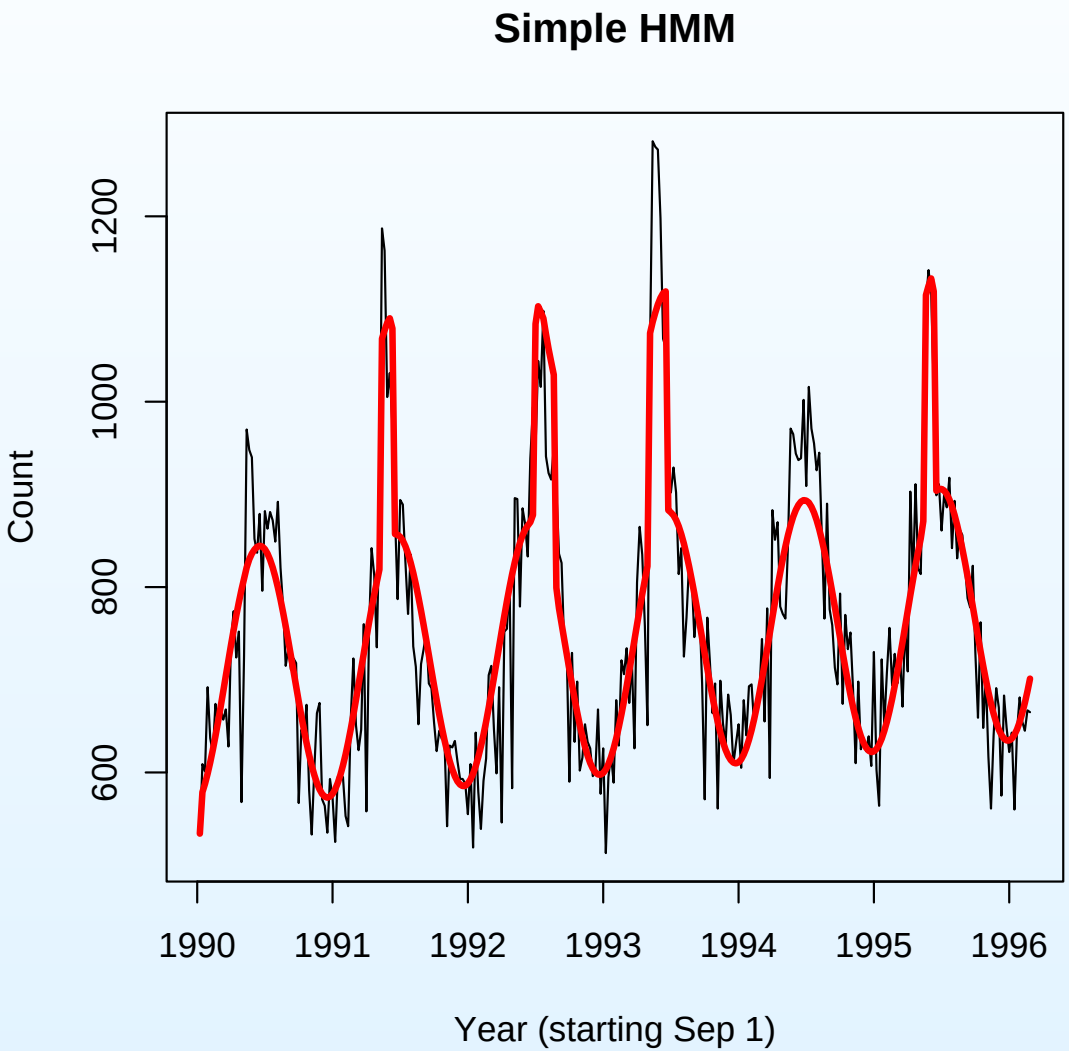
Models

1. Traditional cyclic model (Serfling). OLS regression with terms for intercept, linear trend, two periodic terms for sinusoid with phase shift.
2. Periodic auto-regression (PARMA) with fixed order (1,0) fits cyclic model plus additional ARMA terms.
3. Naive 2-state HMM. Non-epidemic state follows Serfling. Epidemic state modeled with simple mean shift.
4. 2-state AR-HMM. Non-epidemic state, data follow PARMA. Epidemic state auto-regresses deviation from cyclic baseline.

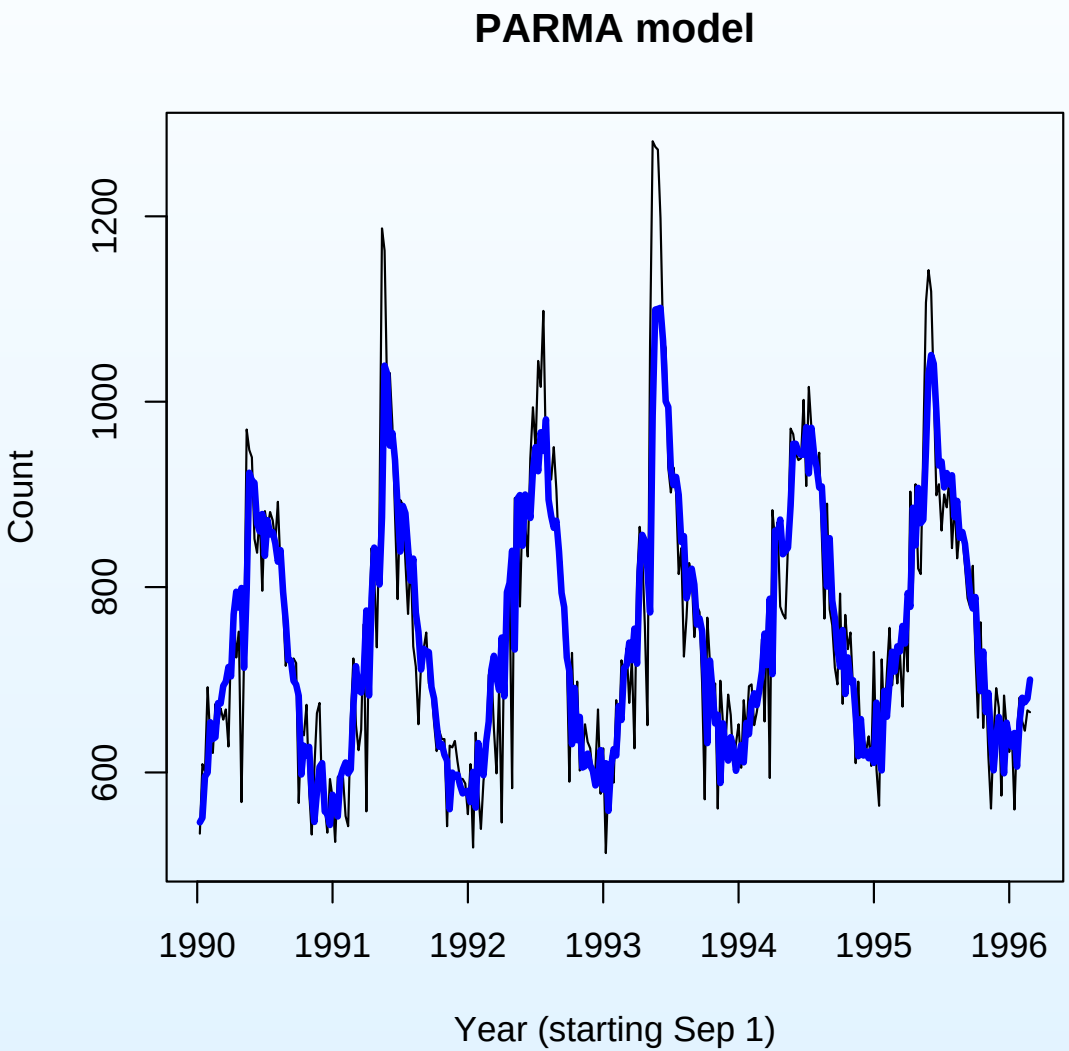
Serfling



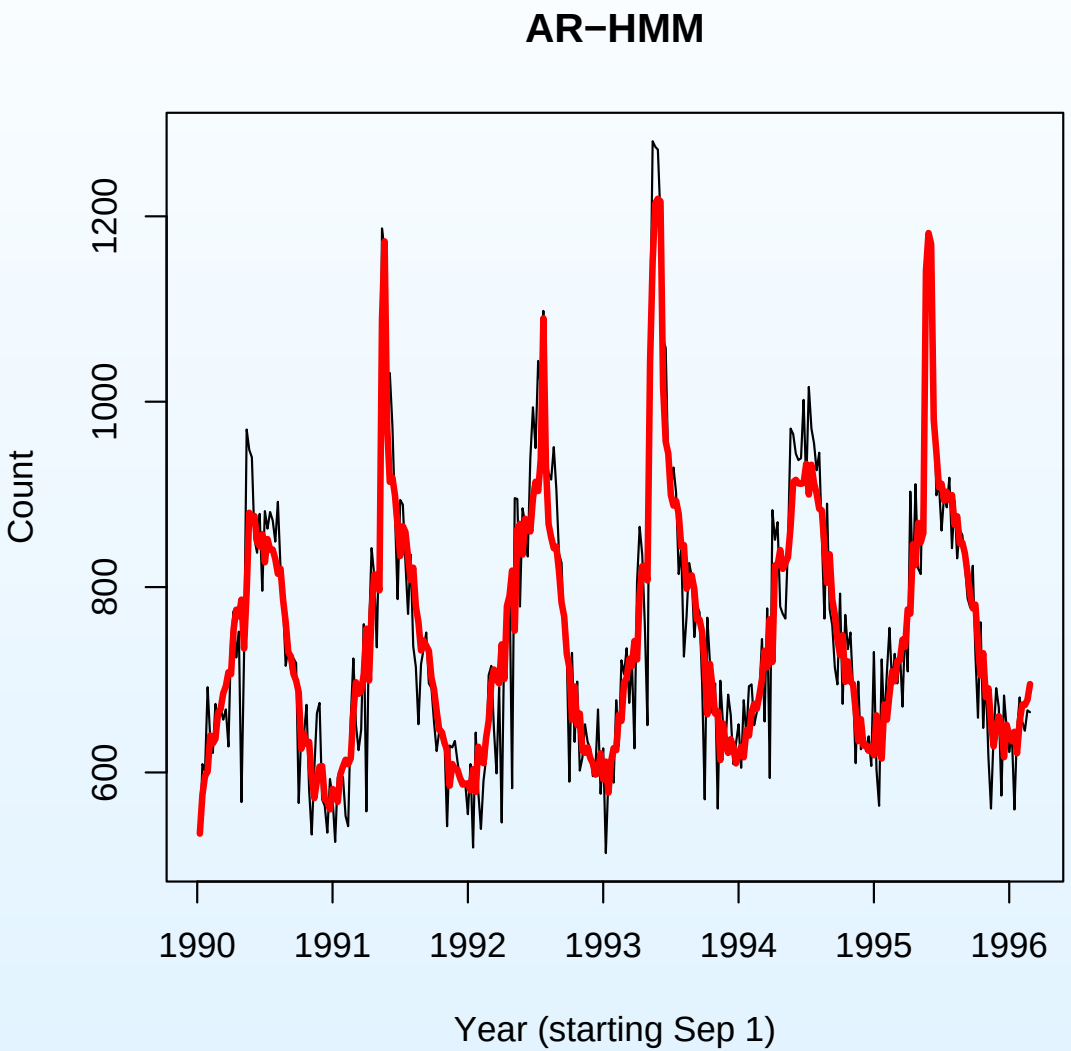
Simple HMM



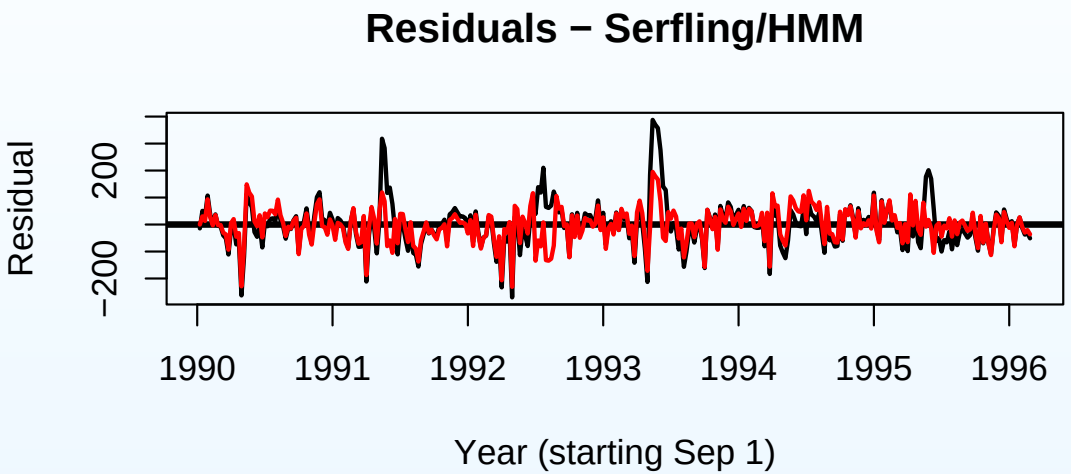
PARMA

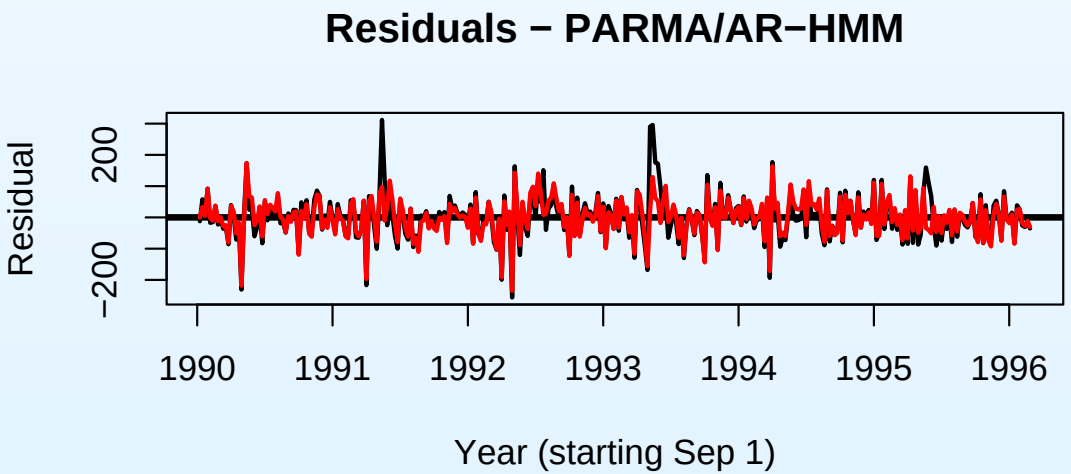


AR-HMM

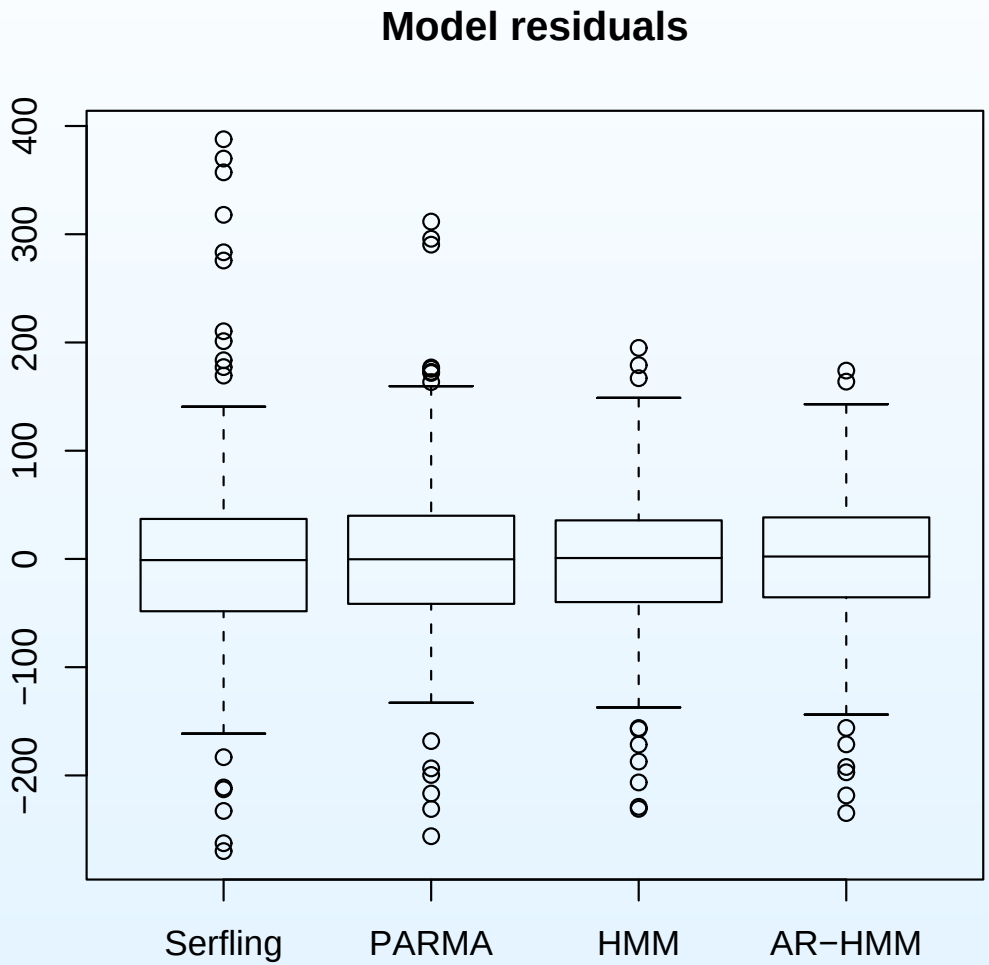


Residuals





Residuals

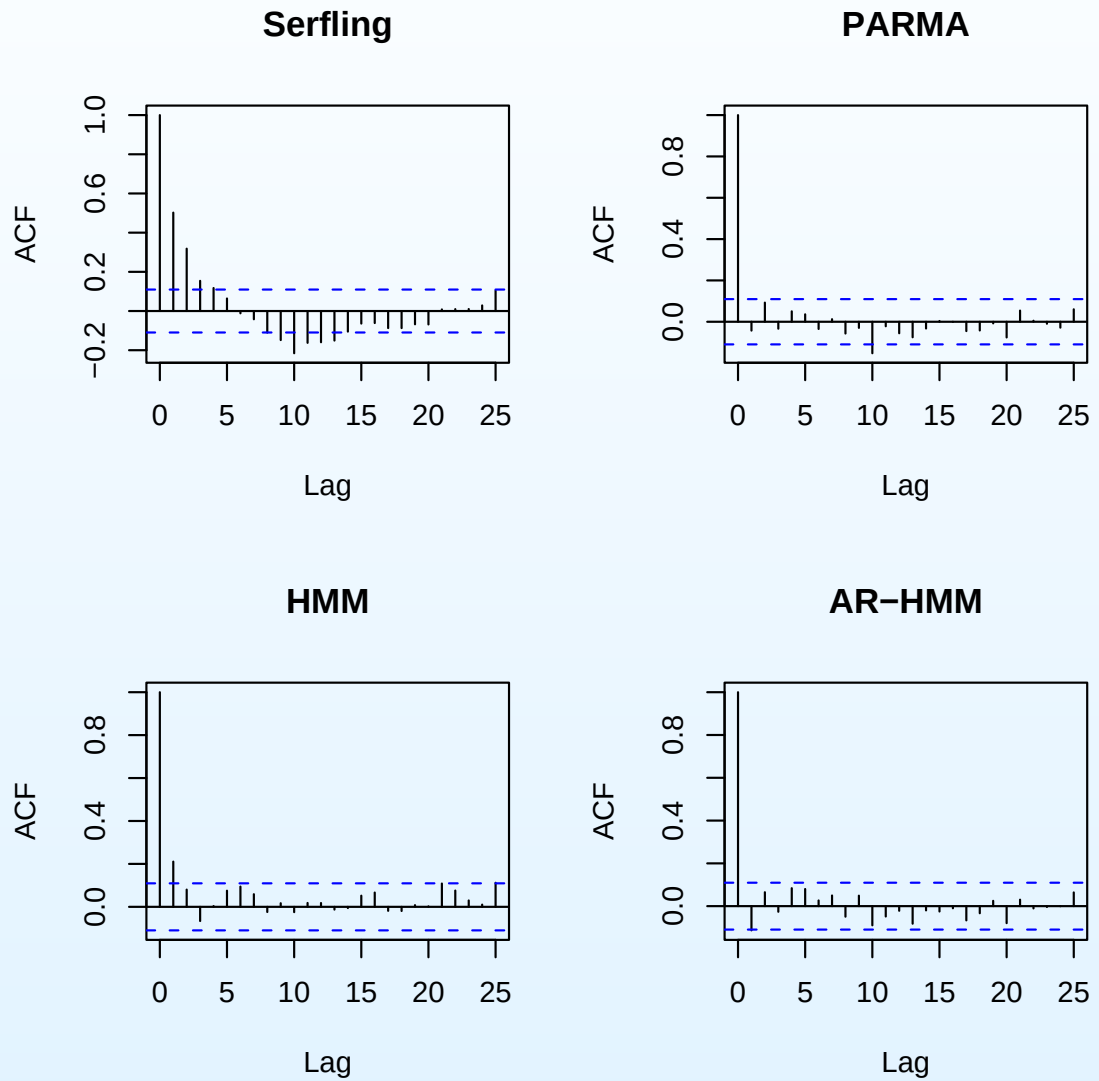


Residuals

Both HMMs provide a roughly 25% reduction in RMSE from Serfling, roughly 10% reduction for PARMA.

Model	RMSE
Serfling	83.3
PARMA	72.0
Simple HMM	63.7
AR-HMM	60.4

ACF of residuals



Conclusions

Preliminary conclusions from model-fitting:

- HMMs offer superior model fit during epidemic periods. AR components do not offer much improvement during non-epidemic period.
- Both models with AR component eliminate auto-correlation of residuals. Important for use of control chart detection methods (e.g. Shewhart, CUSUM).
- Interpretability of latent variable (for two-state models) provides immediate benefit beyond better fit.
- Many-state models ($k > 2$) prove difficult to fit for convergence reasons.

Future work

Integration

Bayesian methodology for integration of multiple time series:

- Developed for gene expression data.
- Bottom-up heuristic search to aggregate time series data; likelihood criterion using model specification to identify “clusters” of time series.
- Hypothesis: cluster assignments will vary over time; possibly dependent on circulating strain, point of origin; less evidence of diffusion in recent years.

Diffusion of influenza

Evidence for diffusion dynamics?

- Standardization of multiple time series to allow for direct comparison across geographic regions.
- Comparison of standardized counts across distances to quantify diffusion over course of surveillance period. Use L^2 norm, cross-correlation, for dissimilarity measure between series.
- Eventual goal: development of true spatio-temporal model for influenza activity.

Acknowledgements

- Many thanks to Nina Fefferman, DIMACS, and working group participants.
- The author gratefully acknowledges the contributions of his co-author Dr. Sebastiani and research assistant Suporn Sukpraprut.
- Thanks also to collaborators at the Harvard School of Public Health: Marcello Pagano, Laura Forsberg, Caroline Jeffery, and Miriam Nuño.
- Willie Anderson of CDC (NCHS) graciously offered assistance in acquiring historical data in electronic format.
- Research partially supported by a pilot grant originating from the Blood Center of Wisconsin, via NIAID grant U19 AI62627-02.