Penalized B-Spline Regression to Analyze Trends in Reported Foodborne Illness

Mark Powell
U.S. Department of Agriculture, Office of Risk Assessment and Cost-Benefit Analysis
Washington, DC

Society for Risk Analysis
6-10 December 2015
Arlington, VA

INTRODUCTION

- Objective: Analyze temporal patterns in reported U.S. foodborne illness without specifying model form
- Illustrative Case: Salmonella
  - Annual and seasonal patterns
  - All Salmonella serotypes, principal serotypes

DATA

- FoodNet (Foodborne Diseases Active Surveillance Network)
- Reported illness counts by site, year, and month
- Population size by site and year
- Site composition stable since 2004

- All Salmonella Serotypes
- Principal Salmonella Serotypes
  - Typhimurium
  - Enteritidis
  - Newport
  - Account for over 40% of serotyped strains

METHODS

- Penalized B-spline Regression
  - Semi-parametric method – no assumed trend model form
  - B-spline basis functions provide local control, local fit is insensitive to points far removed
  - Penalized form of B-spline regression is insensitive to number, placement of join-points (“knots”)
- Wide range of applications

METHODS

- Generalized Additive Model for Poisson Regression
- $\text{Log}(E[y_i]) = \log(\text{population}_i)+\beta_0+f(\text{year}_i)+\varepsilon_i$
  - Smooth $f(\text{year}_i) = \sum B_k(\text{year}_i)\beta_k$
    - $B_k(x)$ = B-spline basis function
  - Year (nx1 vector) $\rightarrow$ **X** (nxk matrix)
    - Fit the model with basis functions as covariates
- $\text{Log}(E[y_{ij}]) = \log(\text{population}_{ij})+\beta_0+f_1(\text{year}_i)+f_2(\text{month}_j)+\varepsilon_{ij}$

METHODS

- At any given point, q+1 B-splines are non-zero (local control)
  - q = B-spline degree (e.g., q=3 for cubic)
  - B-splines sum to 1
- Basis dimension (k) = q + n’ (unconstrained)
  - n’ = no. intervals along domain
  - e.g., 2 internal knots divides domain into n’ = 3 intervals
- Eilers and Marx (1996) provides recursive algorithm for B-spline basis functions for uniformly spaced knots
- In practice, need to impose identifiability constraint → k-1 orthogonal columns (QR decomposition)
- Smoothness controlled by penalty term, fit insensitive to basis dimension

METHODS

- P-IRLS to obtain GLM likelihood maximization, s.t. smooth
- Given λ, min: $\left\|\sqrt{W}(z - X\beta)\right\|^2 + \lambda\beta^T S\beta$
  - λ = curvature penalty parameter
  - $w_i \propto [V(\mu_i)g'(\mu_i)^2]^{-1}$
  - $V[y_i] = \phi\mu_i$ (Generalized Poisson)
  - $z_i = g'(\mu_i)(y_i - \mu_i) + X_i\beta$
  - X = (constrained) design matrix
  - g = link function (log)
  - S (penalty matrix) = $D^T D$
    - penalize differences among neighboring β coefficients
  - For D = second order difference matrix ~ $\int[f''(x)]^2 dx$
    - measure of total curvature

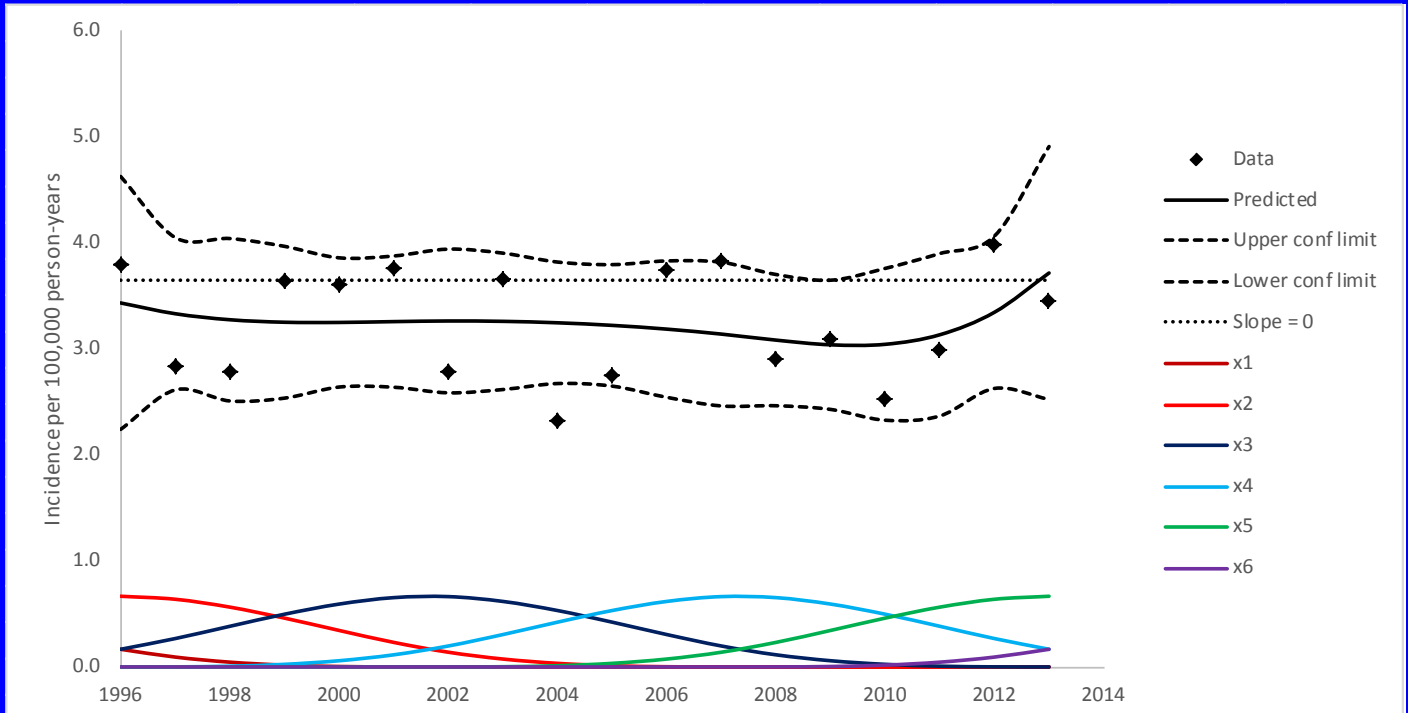
METHODS

- Select degree of smoothness (λ) based on model selection criterion (GCV)
- Effective degrees of freedom (edf) = tr(A)
  - where $\hat{\mu} = Ay$
- With λ = 0, tr(A) = k-1
- As λ→∞, GAM → Log-Linear Model (X →1 edf)

METHODS

- Uniform cubic B-spline basis with 2 internal knots
  - k = q(3)+ n'(3) = 6 unconstrained basis functions
- S (penalty matrix): 2nd order difference matrix
- All Sites
  - Composition of FoodNet sites stable since 2004
- Original 5 Sites
  - Attempt to control for changes in FoodNet composition over time

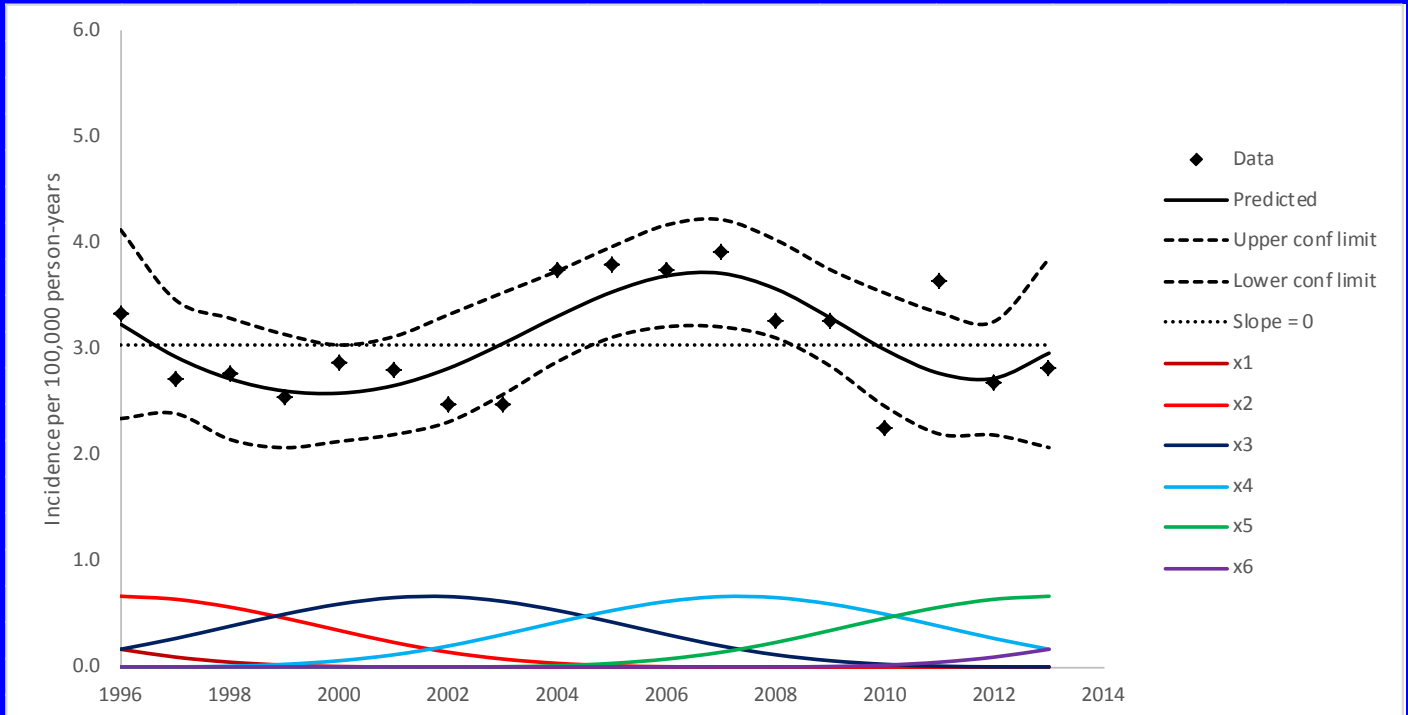
METHODS
B-Spline Provides Flexibility



	PENALTY	0.00										
	AICC	1.351276										

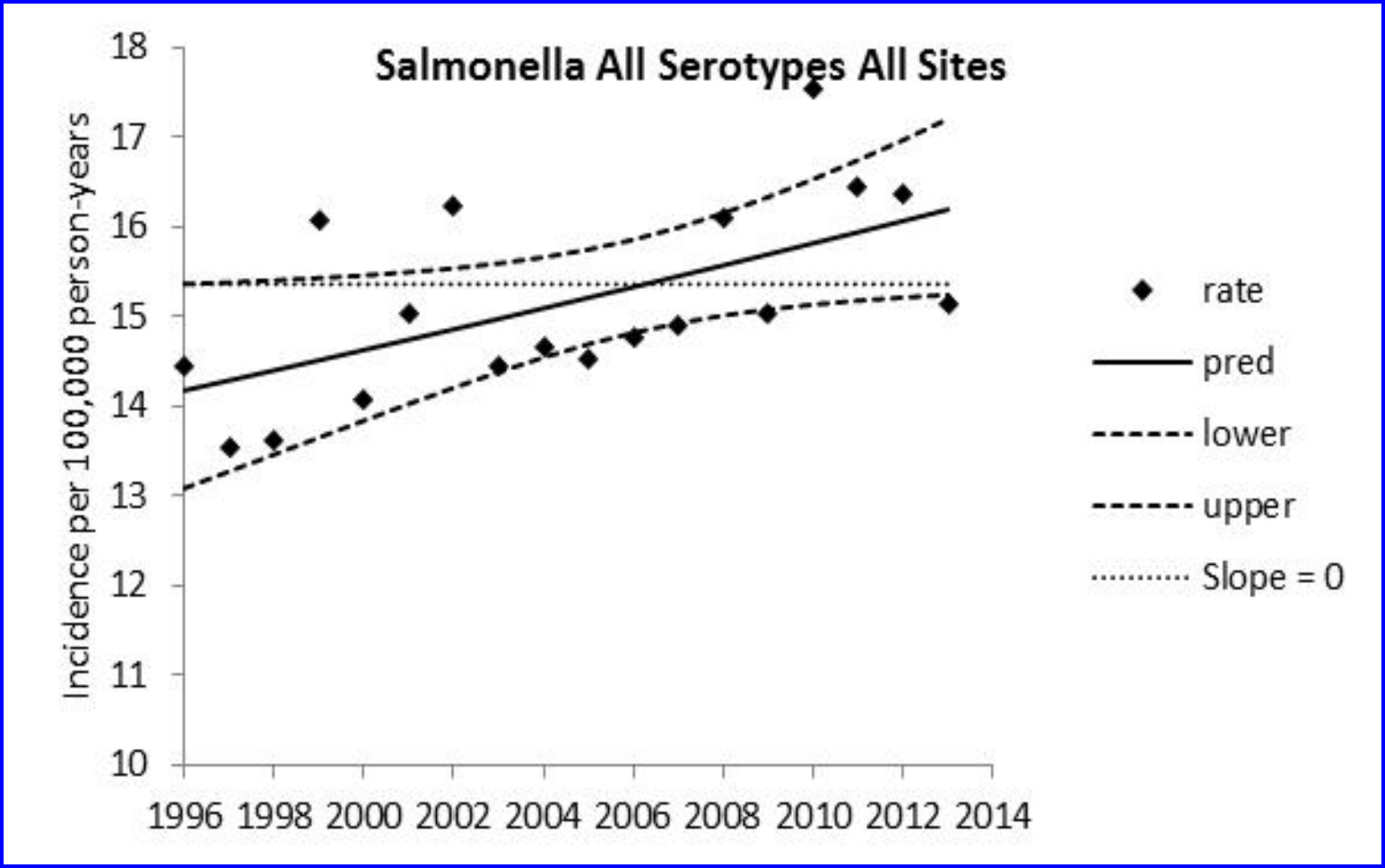
METHODS

Penalized B-Spline Avoids Overfitting



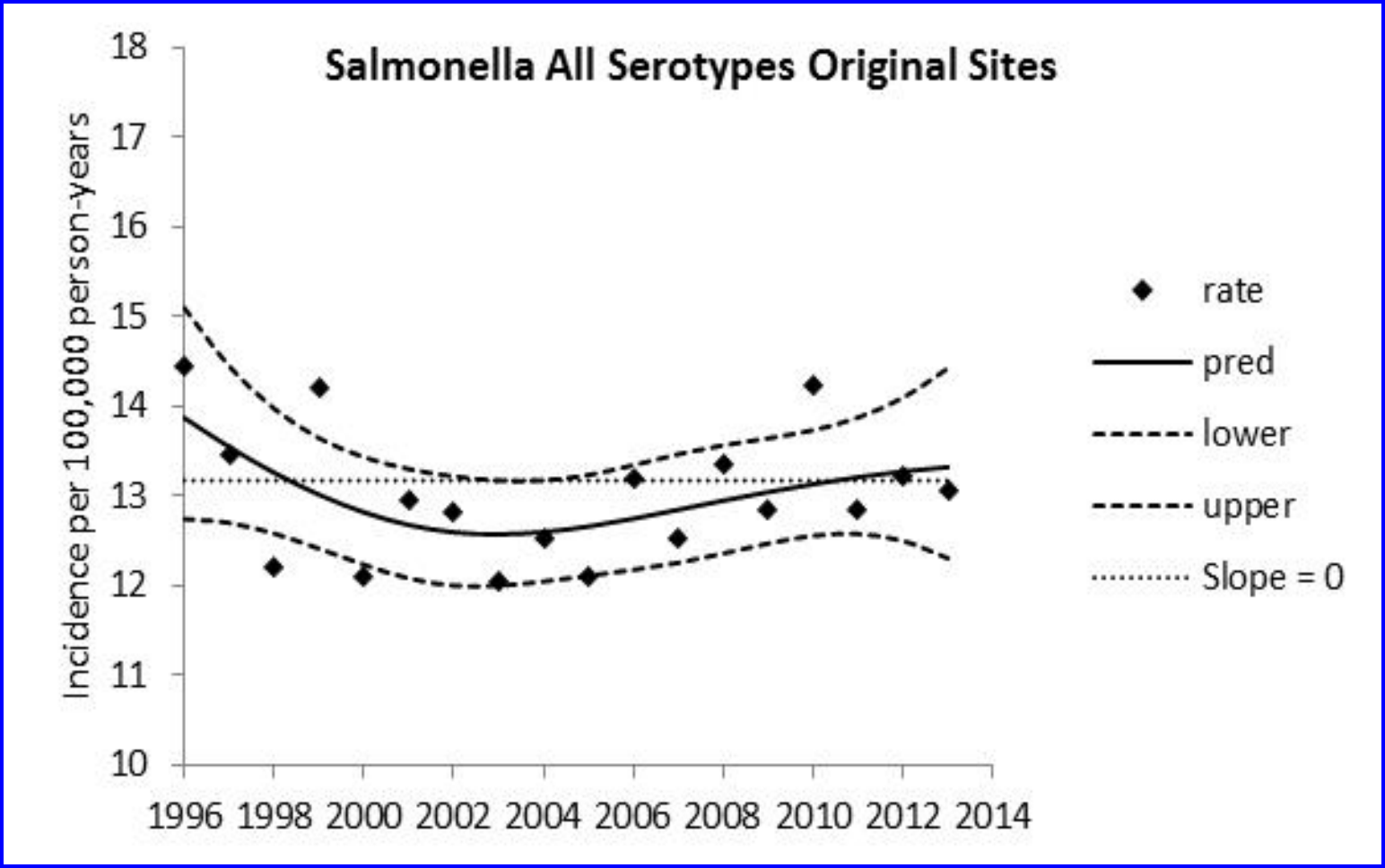
	PENALTY	0.00
	AICC	0.766499

RESULTS



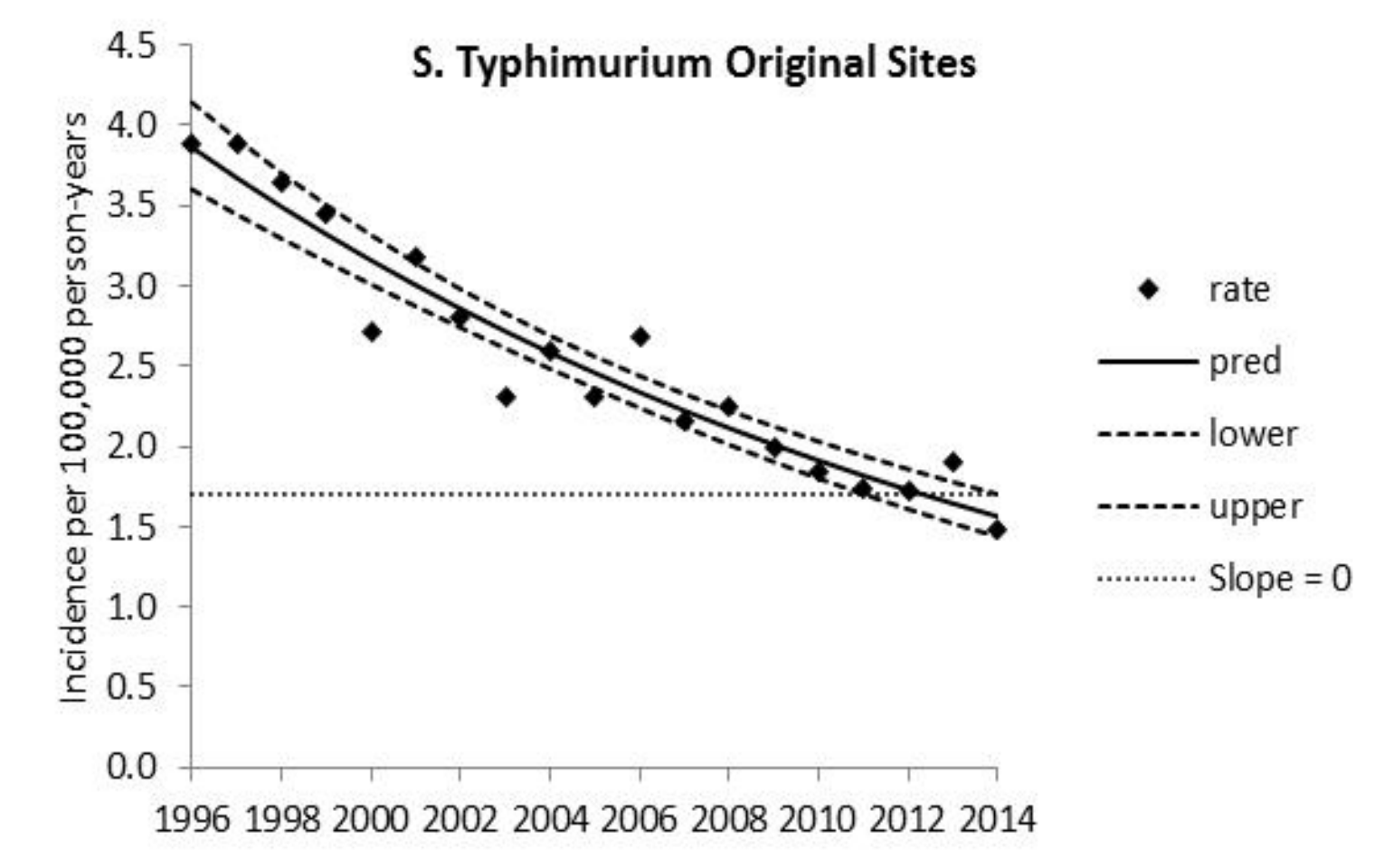
Powell: P-Spline Regression

RESULTS



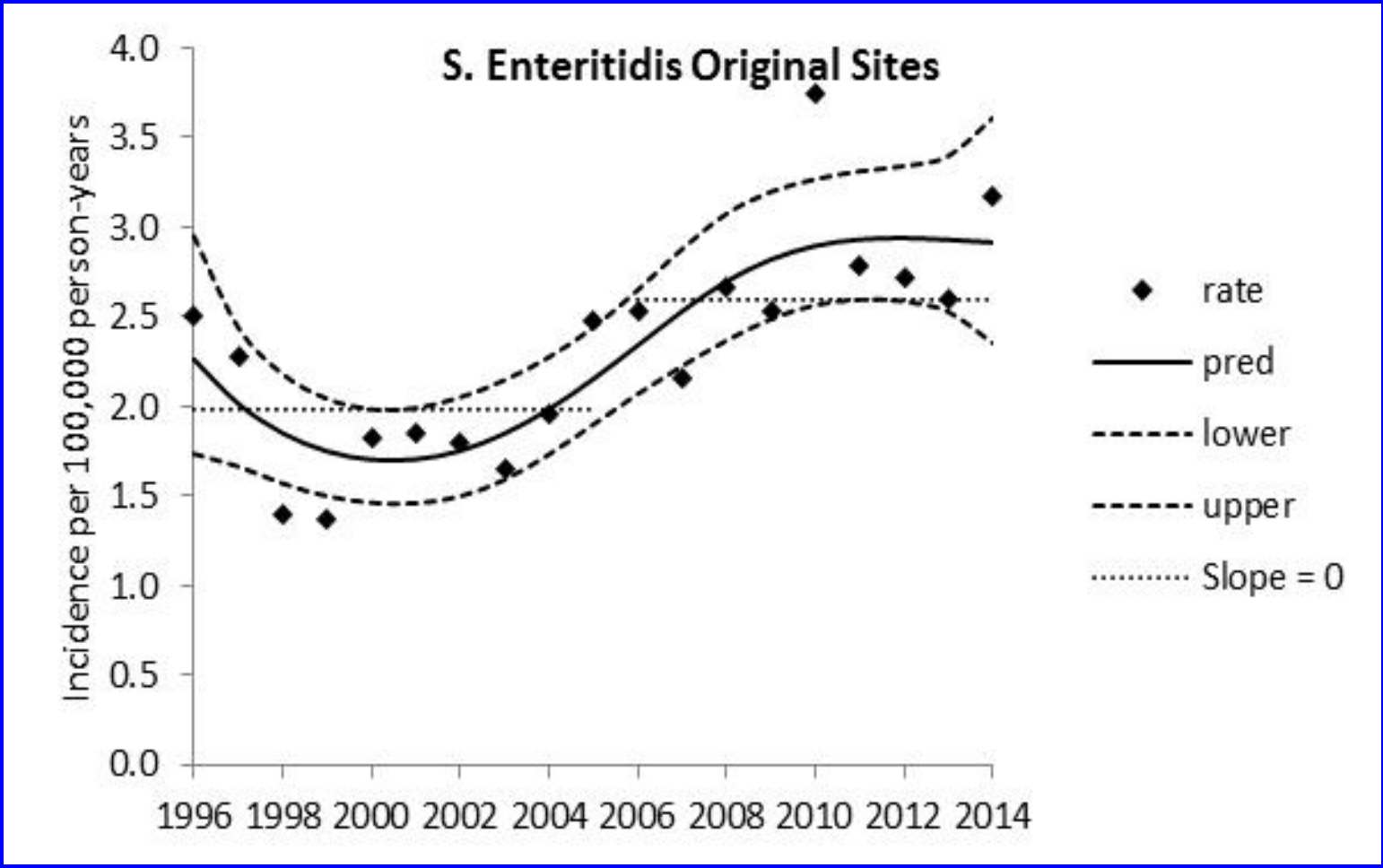
Powell: P-Spline Regression

RESULTS



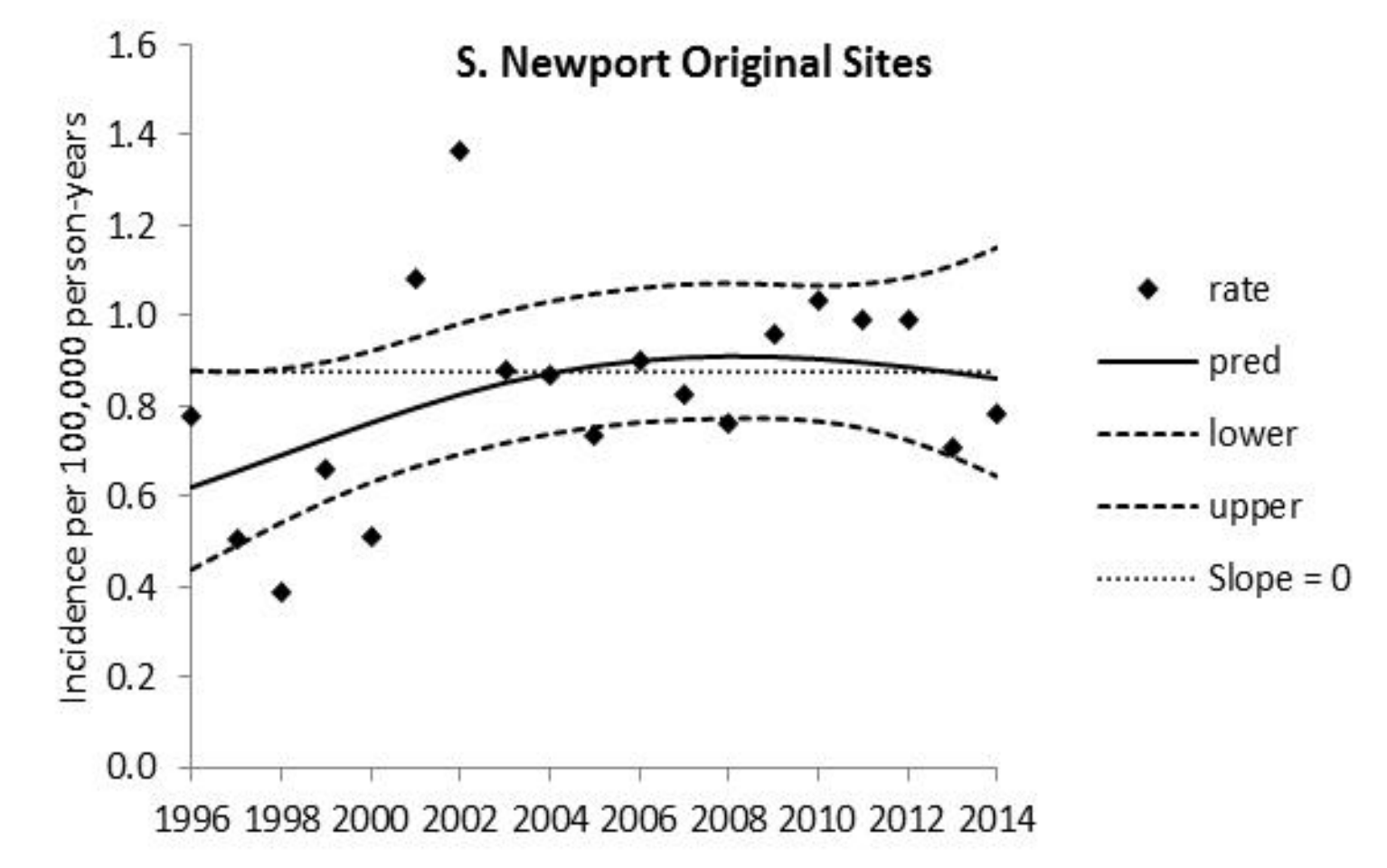
Powell: P-Spline Regression

RESULTS



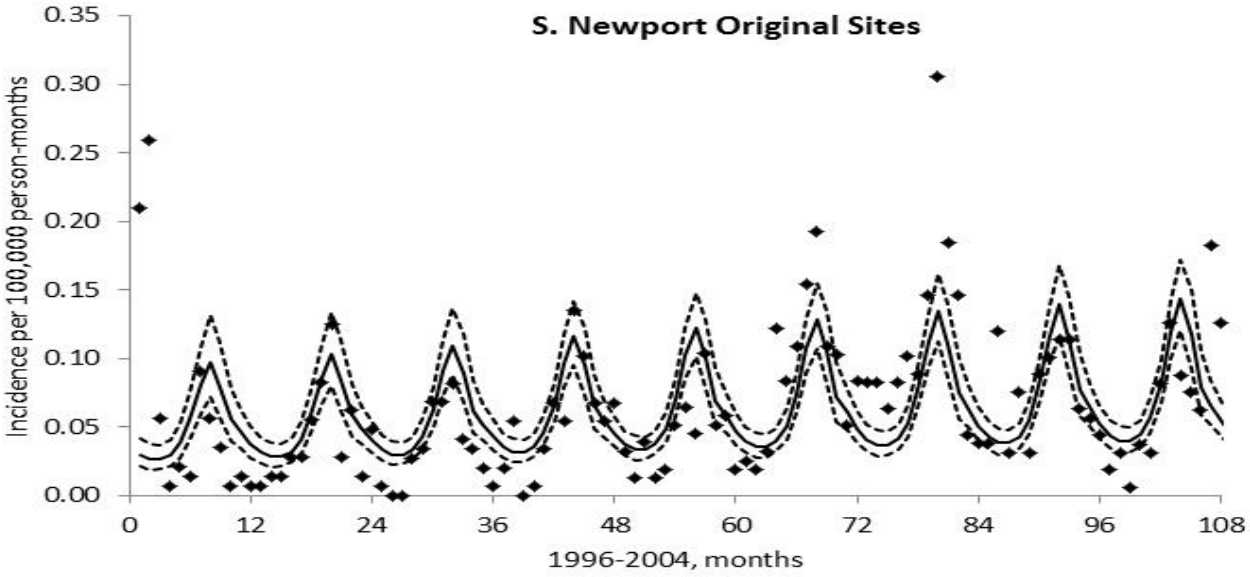
Powell: P-Spline Regression

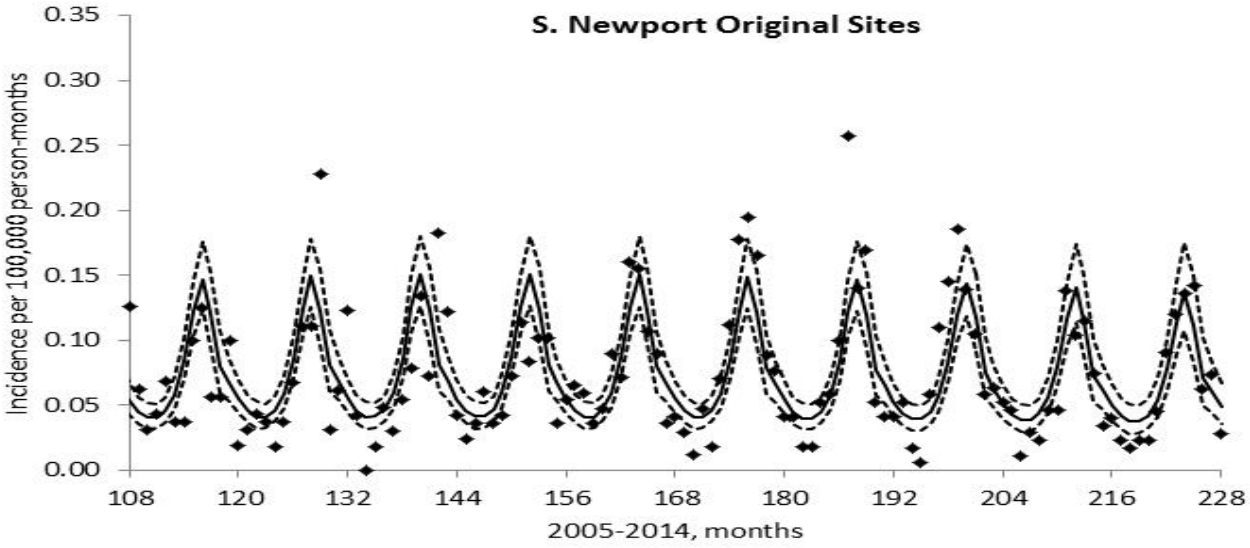
RESULTS



Powell: P-Spline Regression

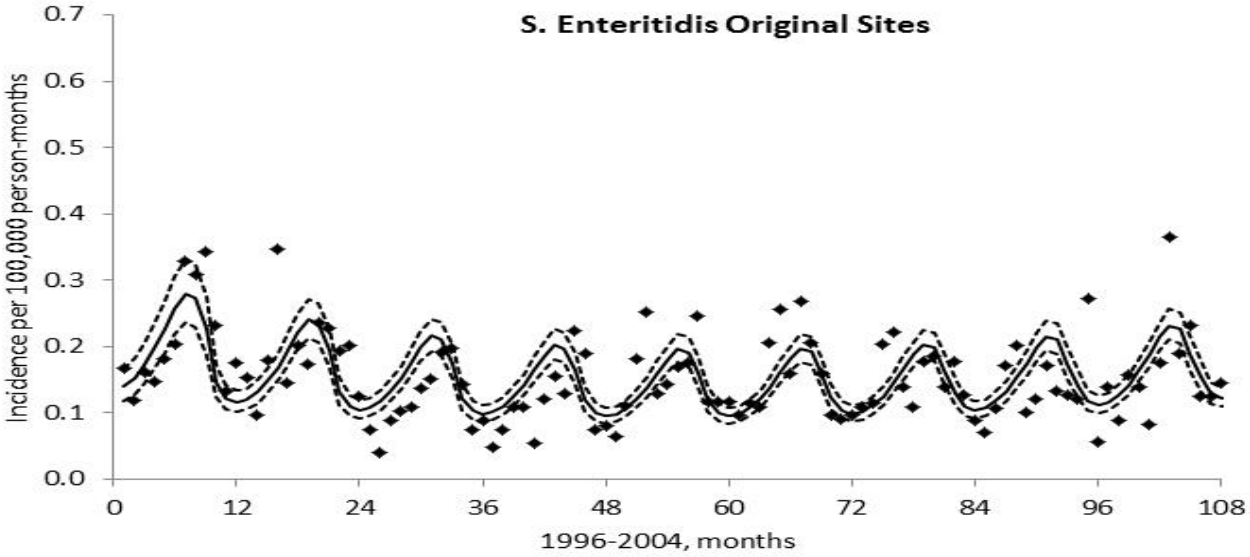
RESULTS

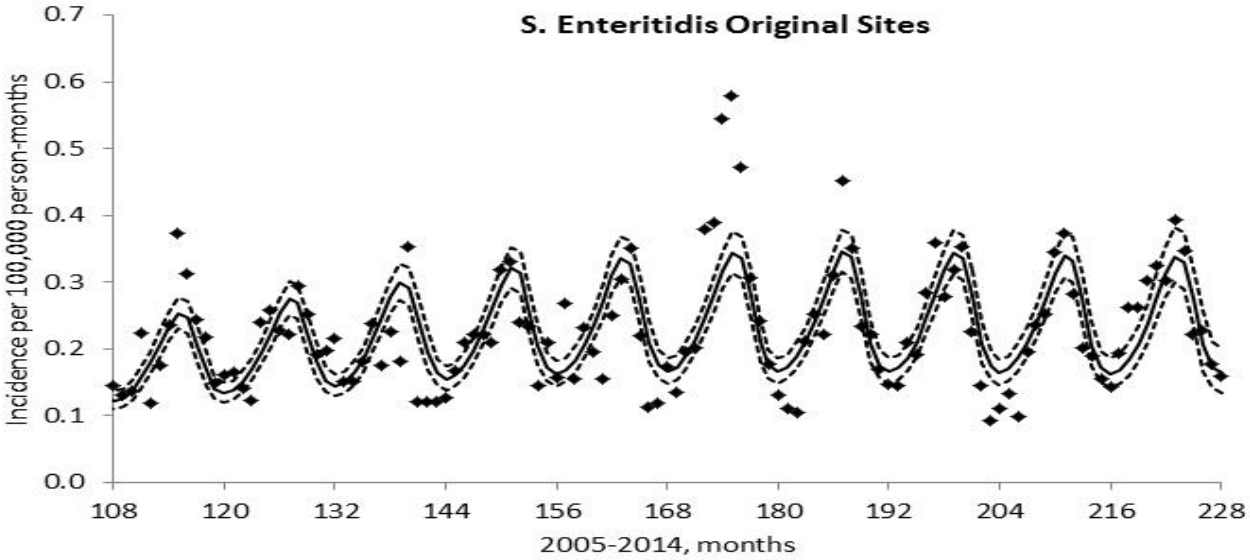




Powell: P-Spline Regression

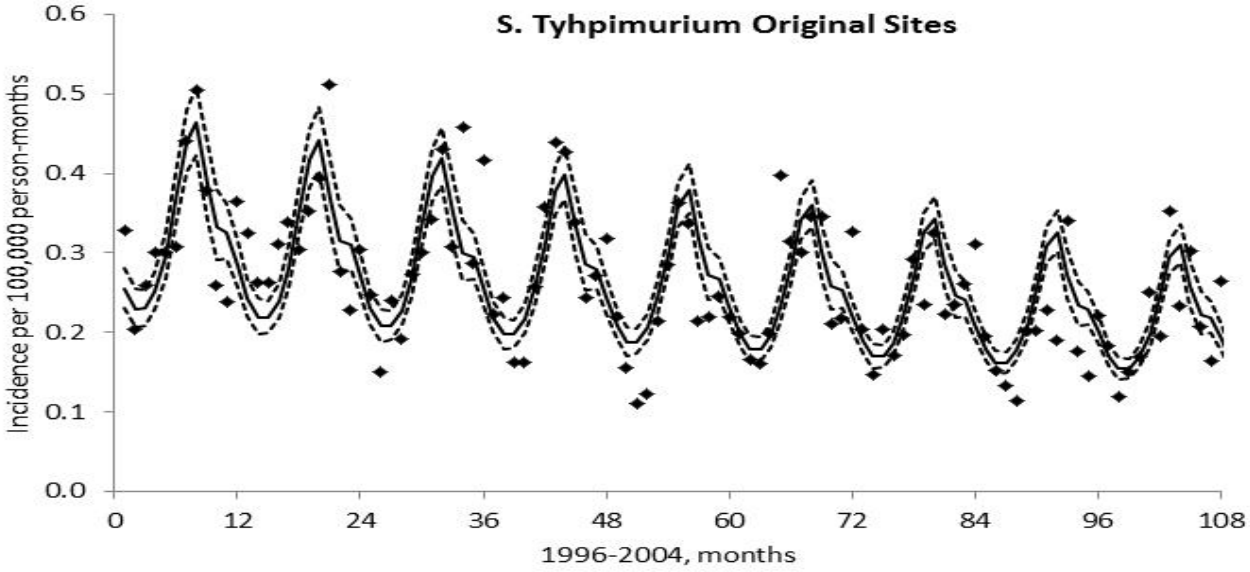
RESULTS

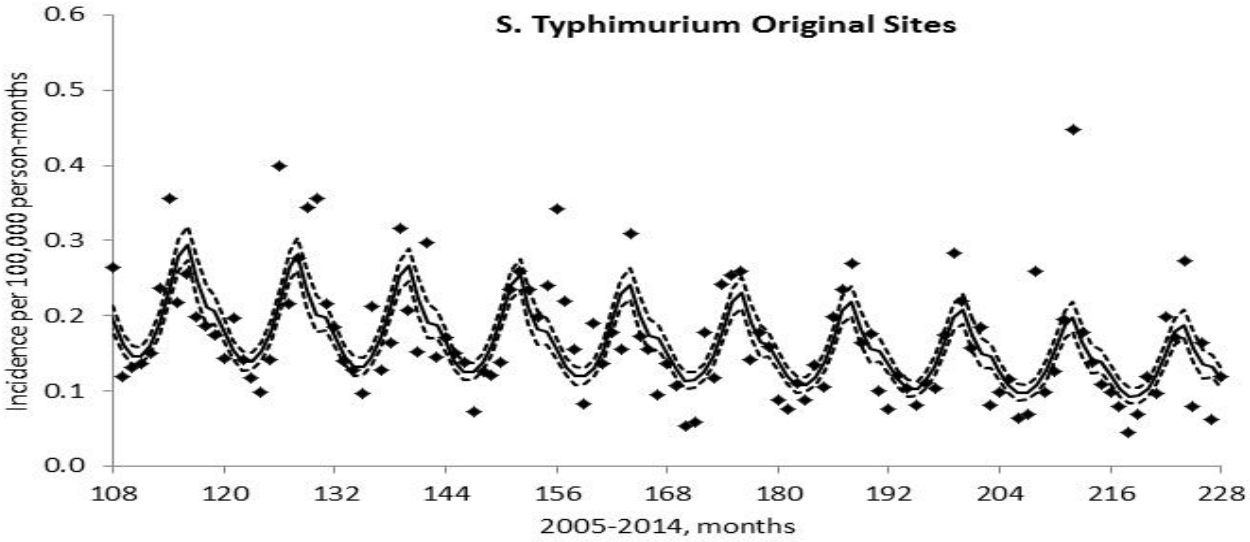




Powell: P-Spline Regression

RESULTS





Powell: P-Spline Regression

Limitations

- Serotype results are preliminary, work in progress
- Less smoothness is imposed at domain boundaries
- Reported illness is a proxy, not true incidence
- Not all FoodNet reported illness is foodborne
- Descriptive model, not infer causes
- Uncertainty about generalizing from FoodNet population to national level not quantified

Acknowledgements

- Data Provided by Foodborne Diseases Active Surveillance Network, CDC
  - Stacy Crim, CDC
  - Mike Hoekstra, CDC
  - Weidong Gu, CDC
  - Mike Williams, FSIS

Disclaimers

- The opinions expressed herein are the views of the author and do not necessarily reflect the official policy or position of the United States Department of Agriculture. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government.
- The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.