Case Studies Using Credibility and Corrected Adaptively Truncated Likelihood Methods

Vytaras Brazauskas [a,b]

University of Wisconsin-Milwaukee

CAS Centennial Celebration and Annual Meeting

New York, NY, November 9–12, 2014

[a] Joint work with Harald Dornheim (KPMG Switzerland)

[b] Supported by a grant from the Casualty Actuarial Society

Outline

1. **Introduction**

    ∗ Credibility
    ∗ Robustness

2. **CATL Credibility**

    ∗ CATL Procedure
    ∗ Robust Ratemaking

3. **Case Studies**

    ∗ Bodily Injury Data
    ∗ Medicare Data
    ∗ Real Estate Data

4. **Final Remarks**

1. Introduction

Credibility

- **Idea**

$$\text{Premium} \;=\; Z \times \text{Individual Experience} \;+\; (1 - Z) \times \text{Portfolio Rate}$$

- **History**
  - ∗ MOWBRAY (1914), WHITNEY (1918):
    Limited fluctuation credibility theory; premium stability.
  - ∗ BÜHLMANN (1967), BÜHLMANN, STRAUB (1970):
    Greatest accuracy credibility theory; Bayesian statistics.
  - ∗ HACHEMEISTER (1975):
    Credibility model linked to regression; bodily injury data.

- **Recent Developments**
  - ∗ FREES, YOUNG, LUO (1999, 2001):
    Longitudinal data analysis interpretation; case studies.
  - ∗ FELLINGHAM, TOLLEY, HERZOG (2005):
    Claims experience for health insurance coverages in IL and WI.
  - ∗ GUSZCZA (2008):
    Introduction to hierarchical models; loss reserving exercise.
  - ∗ KLINKER (2011):
    Generalized LMM; ISO data; Bühlmann-Straub credibility.
- **Problems**
  - ∗ PITSELIS (2002, 2008), DORNHEIM, BRAZAUSKAS (2007):
    Structural parameters; (RE)ML estimation; (non)robustness.

Robustness

- **Idea**
  - ∗ MODEL RISK (model misspecification; wrong assumptions)
  - ∗ DATA QUALITY (measurement errors; outliers; typos)
- **Objectives**
  - ∗ Focus on parametric models, their fitting to the observed data, and identification of outliers.
  - ∗ Construct methods with limited sensitivity to changes in the underlying assumptions and to “unexpected” data points (i.e., robust methods).
- **Primary Tool**
  - ∗ INFLUENCE FUNCTION (directional derivative; helps to quantify the method’s robustness and efficiency; two competing criteria)

- **Typical Solutions**
  - ∗ L-STATISTICS (linear combinations of order statistics)
  - ∗ M-STATISTICS (maximum likelihood type statistics)
  - ∗ R-STATISTICS (statistics based on ranks)
- **Impact**
  - ∗ Humble beginnings (location parameter of bell-shaped curve)
  - ∗ Steady growth (generalized linear models; time series analysis)
  - ∗ Global reach (sciences; engineering; insurance; finance)
- **Take-Home Lessons**
  - ∗ Robust statistics is about model risk management!
  - ∗ Robust methods add most value to the modeling process when the underlying model has many unknown parameters and assumptions!

2. CATL Credibility

CATL Procedure

- **Scope**

  ∗ Corrected Adaptively Truncated Likelihood (CATL) methods designed for situations when claims approximately follow a fat-tailed distribution.

- **History**

  ∗ Introduced and developed by Dornheim, Brazauskas (2011a,b); Builds on the work of Rousseeuw, Leroy (1987) and Gervini, Yohai (2002).

- **Special Features**

  ∗ Protection against within-risk and between-risk outliers.

  ∗ Automatic identification and removal of atypical data points; no expert opinion; no graphical tools; no data-specific predictor variables.

- **Three-Step Procedure**

1. WITHIN-RISK Outliers

$$\big(\mathbf{x}_{i1}, \mathbf{z}_{i1}, \log(y_{i1}), \upsilon_{i1}\big), \ldots, \big(\mathbf{x}_{i\tau_i}, \mathbf{z}_{i\tau_i}, \log(y_{i\tau_i}), \upsilon_{i\tau_i}\big), \quad i = 1, \ldots, I$$

Apply highly robust estimators for location and scale parameters, and remove observations whose standardized residuals are extreme:

$$\big(\mathbf{x}^*_{i1}, \mathbf{z}^*_{i1}, \log(y^*_{i1}), \upsilon^*_{i1}\big), \ldots, \big(\mathbf{x}^*_{i\tau^*_i}, \mathbf{z}^*_{i\tau^*_i}, \log(y^*_{i\tau^*_i}), \upsilon^*_{i\tau^*_i}\big), \quad i = 1, \ldots, I$$

2. BETWEEN-RISK Outliers

Search the pre-cleaned sample (marked with *) and discard entire risks whose robustified Mahalanobis distance is extreme:

$$\big(\mathbf{x}^{**}_{i1}, \mathbf{z}^{**}_{i1}, \log(y^{**}_{i1}), \upsilon^{**}_{i1}\big), \ldots, \big(\mathbf{x}^{**}_{i\tau^*_i}, \mathbf{z}^{**}_{i\tau^*_i}, \log(y^{**}_{i\tau^*_i}), \upsilon^{**}_{i\tau^*_i}\big), \quad i = 1, \ldots, I^*$$

3. CATL Estimators

Apply likelihood-based methods on the cleaned sample (marked with **):

$$\widehat{\boldsymbol{\beta}}_{\text{CATL}}, \quad \widehat{\boldsymbol{\theta}}_{\text{CATL}} = (\widehat{\sigma}^2_{\alpha_1}, \ldots, \widehat{\sigma}^2_{\alpha_q}, \widehat{\sigma}^2_{\varepsilon}) \text{ and } \widehat{\boldsymbol{\alpha}}_{\text{rBLUP}, \, i},$$

$$\widehat{\boldsymbol{\lambda}}_i = \mathbf{X}^{**}_i \widehat{\boldsymbol{\beta}}_{\text{CATL}} + \mathbf{Z}^{**}_i \widehat{\boldsymbol{\alpha}}_{\text{rBLUP}, \, i} + \widehat{\mathbf{E}}_{F_0}(\boldsymbol{\varepsilon}_i), \quad i = 1, \ldots, I$$

Robust Ratemaking

- **Ordinary Premiums**

$$\widehat{\mu}_{it}^{\text{ordinary}} = \widehat{\mu}_{it}^{\text{ordinary}}(\widehat{\boldsymbol{\alpha}}_{\text{rBLUP},i}), \quad t = 1, \ldots, \tau_i + 1, \;\; i = 1, \ldots, I$$

  Limited Expected Value (LEV) of the fitted log-location-scale distribution of claims. The percentile levels of the lower bound q_l and the upper bound q_g used in LEV computations are extreme: 0.1% for q_l and 99.9% for q_g.

- **Excess Claims**

$$\widehat{O}_{it} = \begin{cases} -\widehat{\mu}_{it}^{\text{ordinary}}, & \text{for} \;\; y_{it} < q_l. \\ (y_{it} - q_l) - \widehat{\mu}_{it}^{\text{ordinary}}, & \text{for} \;\; q_l \leq y_{it} < q_g. \\ (q_g - q_l) - \widehat{\mu}_{it}^{\text{ordinary}}, & \text{for} \;\; y_{it} \geq q_g. \end{cases}$$

- **Extraordinary Premium**

  Let I_t denote the number of insureds in the portfolio at time t and let $T = \max_{1 \le i \le I} \tau_i$, the maximum horizon among all risks. For each period $t = 1, \ldots, T$, find the mean cross-sectional overshot of excess claims $\widehat{O}_{\bullet t} = I_t^{-1} \sum_{i=1}^{I_t} \widehat{O}_{it}$, and fit robustly the model

  $$\widehat{O}_{\bullet t} = \mathbf{o}_t \boldsymbol{\xi} + \tilde{\varepsilon}_t, \qquad t = 1, \ldots, T,$$

  where $\mathbf{o}_t$ ($\mathbf{o}_t = 1$) is the vector of covariates for the hypothetical mean of overshots $\boldsymbol{\xi}$. The premium for extraordinary claims is common to all risks:

  $$\mu_{it}^{\text{extra}} = \mathbf{o}_t \widehat{\boldsymbol{\xi}}.$$

- **Final Premiums**

  Portfolio-unbiased CATL credibility premiums:

  $$\widehat{\mu}_{i,\tau_i+1}^{\text{CATL}}(\widehat{\boldsymbol{\alpha}}_{\text{rBLUP},i}) = \widehat{\mu}_{i,\tau_i+1}^{\text{ordinary}}(\widehat{\boldsymbol{\alpha}}_{\text{rBLUP},i}) + \mu_{i,\tau_i+1}^{\text{extra}}, \qquad i = 1, \ldots, I.$$

3. Case Studies

Bodily Injury Data

- **Data**
  - ∗ 12 periods (from 3rd quarter in 1970 to 2nd quarter in 1973) of claims for bodily injury that are covered by a private passenger auto insurance; claims classified by state (5 states with different inflation trends).
  - ∗ RESPONSE VARIABLE: average loss per claim.
- **Model**
  - ∗ LINEAR TREND MODEL: $\mathbf{x}_{it} = \mathbf{z}_{it} = (1, t)'$. The first credibility model linked to regression, introduced by Hachemeister (1975).
- **Objectives**
  - ∗ Model fitting; next-period predictions; sensitivity analysis.

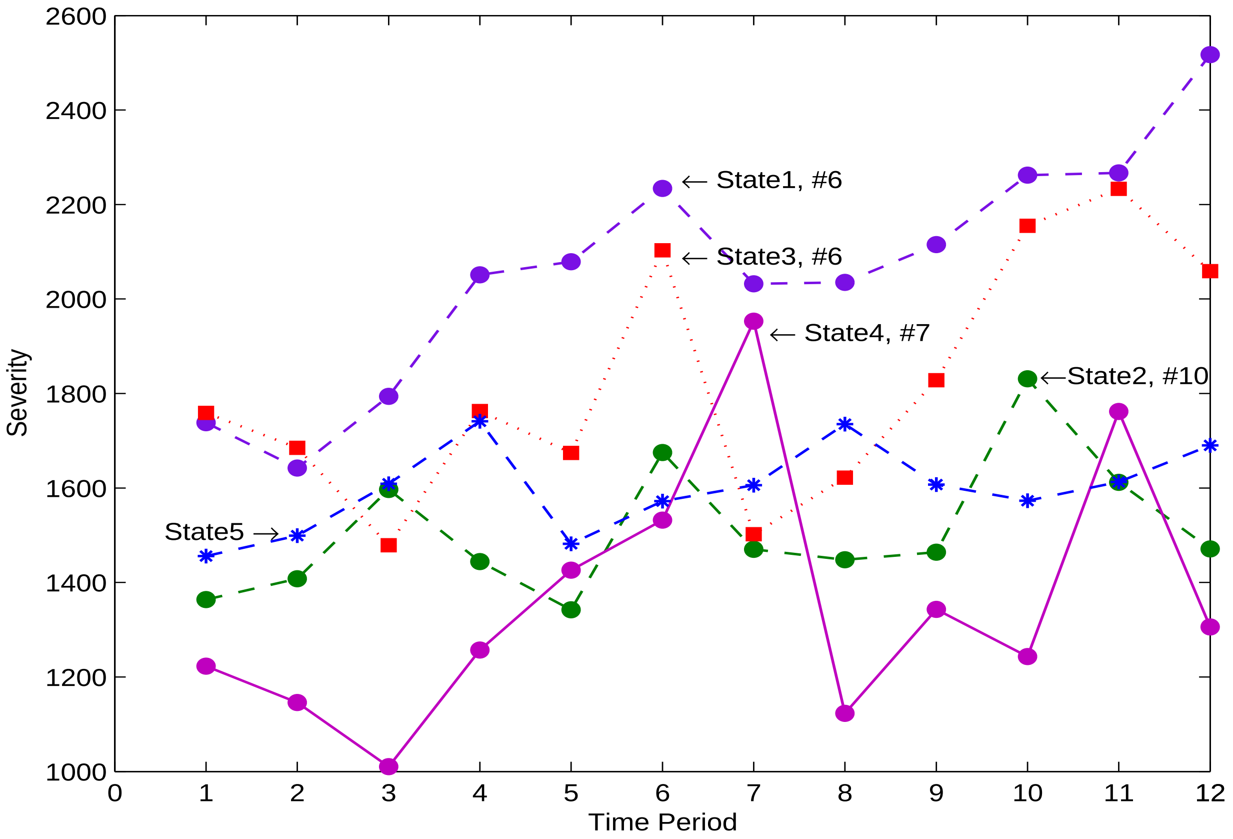

FIGURE 1: Multiple time series plot of the average loss per claim.

TABLE 1: Number of claims per period, v_{it}.

Period	*State*				
	1	*2*	*3*	*4*	*5*
1	7861	1622	1147	407	2902
2	9251	1742	1357	396	3172
3	8706	1523	1329	348	3046
4	8575	1515	1204	341	3068
5	7917	1622	998	315	2693
6	8263	1602	1077	328	2910
7	9456	1964	1277	352	3275
8	8003	1515	1218	331	2697
9	7365	1527	896	287	2663
10	7832	1748	1003	384	3017
11	7849	1654	1108	321	3242
12	9077	1861	1121	342	3425

- **Model Fitting**

  - BASE: Linear trend model (Goovaerts, Hoogstad, 1987).
  - M-RC: Robust credibility based on M-estimators (Pitselis, 2008).
  - MM-RC, GM-RC: Robust credibility based on Multiple M-estimators and Generalized M-estimators (Pitselis, 2002).
  - REML, CATL: REstricted Maximum Likelihood and CATL fitting of the Hachemeister's model (Dornheim, Brazauskas, 2011a,b).
  - REML*, CATL*: REstricted Maximum Likelihood and CATL fitting of the revised Hachemeister's model (Dornheim, Brazauskas, 2011a,b).

NOTE: The regression credibility line must stay between the individual line and the portfolio line! In the revised Hachemeister's model, we take the intercept of the regression line at the center of gravity of the time variable, instead of the origin of the time axis (Bühlmann, Gisler, 1997).

TABLE 2: Next-period predictions for bodily injury data (estimated standard errors provided in parentheses).

Fitting	*Prediction for State*				
Procedure	*1*	*2*	*3*	*4*	*5*
BASE	2436	1650	2073	1507	1759
M-RC	2437	1650	2073	1507	1759
GM-RC	2427	1648	2092	1505	1737
MM-RC	2427	1648	2092	1505	1737
REML	2465 (109)	1625 (122)	2077 (193)	1519 (248)	1695 (77)
CATL	2471 (111)	1545 (74)	2065 (194)	1447 (174)	1691 (57)
REML*	2451 (109)	1661 (123)	2065 (193)	1613 (242)	1706 (78)
CATL*	2450 (113)	1552 (74)	2049 (195)	1477 (172)	1693 (57)

OUTLIERS: State 2 ($t = 6$, $t = 10$), State 4 ($t = 7$), State 5 ($t = 4$).

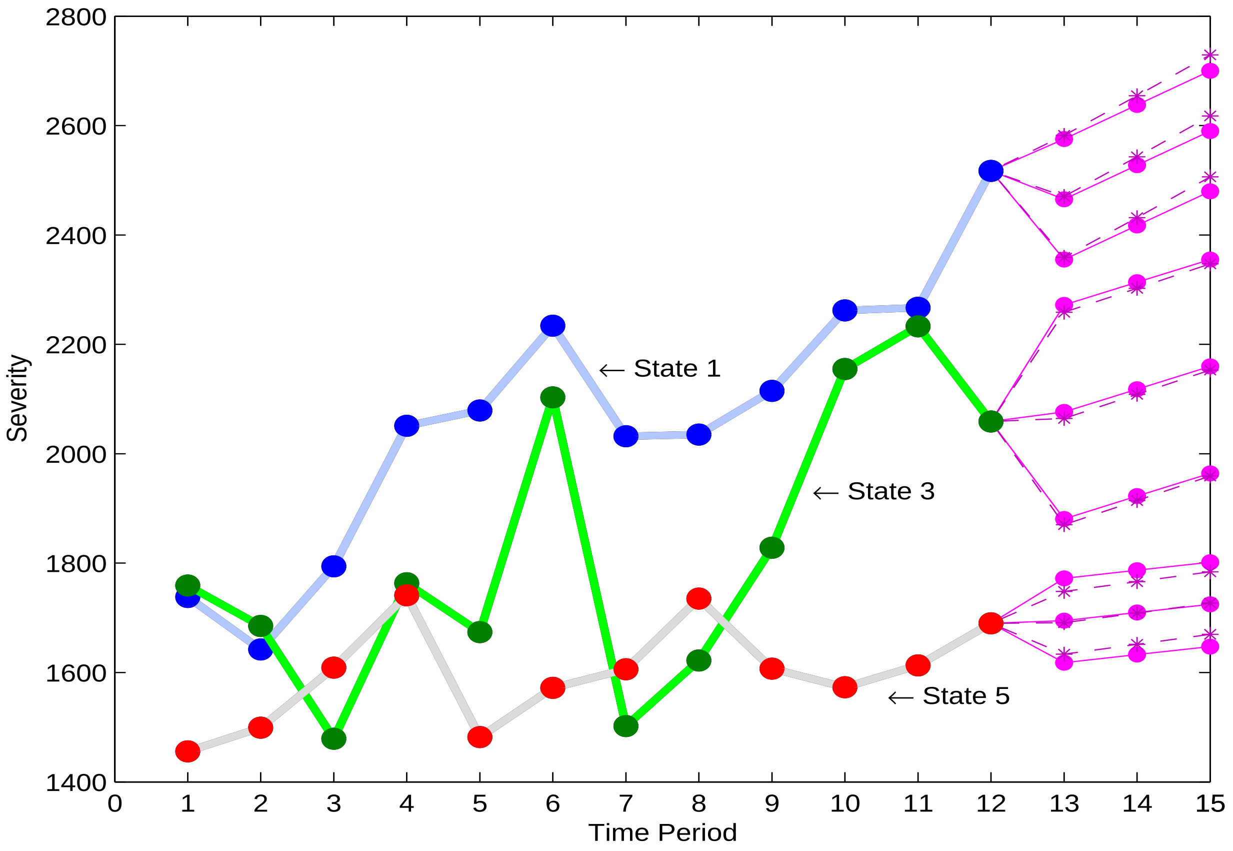

FIGURE 2: One-, two-, three-step predictions for States 1, 3, 5. (CATL: dashed line, marked by $*$. REML: solid line, marked by •.)

TABLE 3: Next-period predictions for contaminated bodily injury data (estimated standard errors provided in parentheses).

Fitting Procedure	*Prediction for State* 1	2	3	4	5
BASE	2501	1826	2181	1994	2596
M-RC	2755	1979	2396	1841	2121
GM-RC	2645	1868	2311	1723	1964
MM-RC	2649	1870	2315	1724	1943
REML	2517 (119)	1852 (150)	2206 (204)	1987 (255)	2542 (829)
CATL	2477 (111)	1550 (74)	2071 (194)	1452 (174)	1689 (60)
REML*	2455 (110)	1949 (166)	2229 (204)	2141 (275)	2629 (818)
CATL*	2459 (112)	1559 (74)	2057 (195)	1484 (172)	1694 (60)

Portfolio contamination: In State 5, the last observation, 1690, is replaced by 5000.

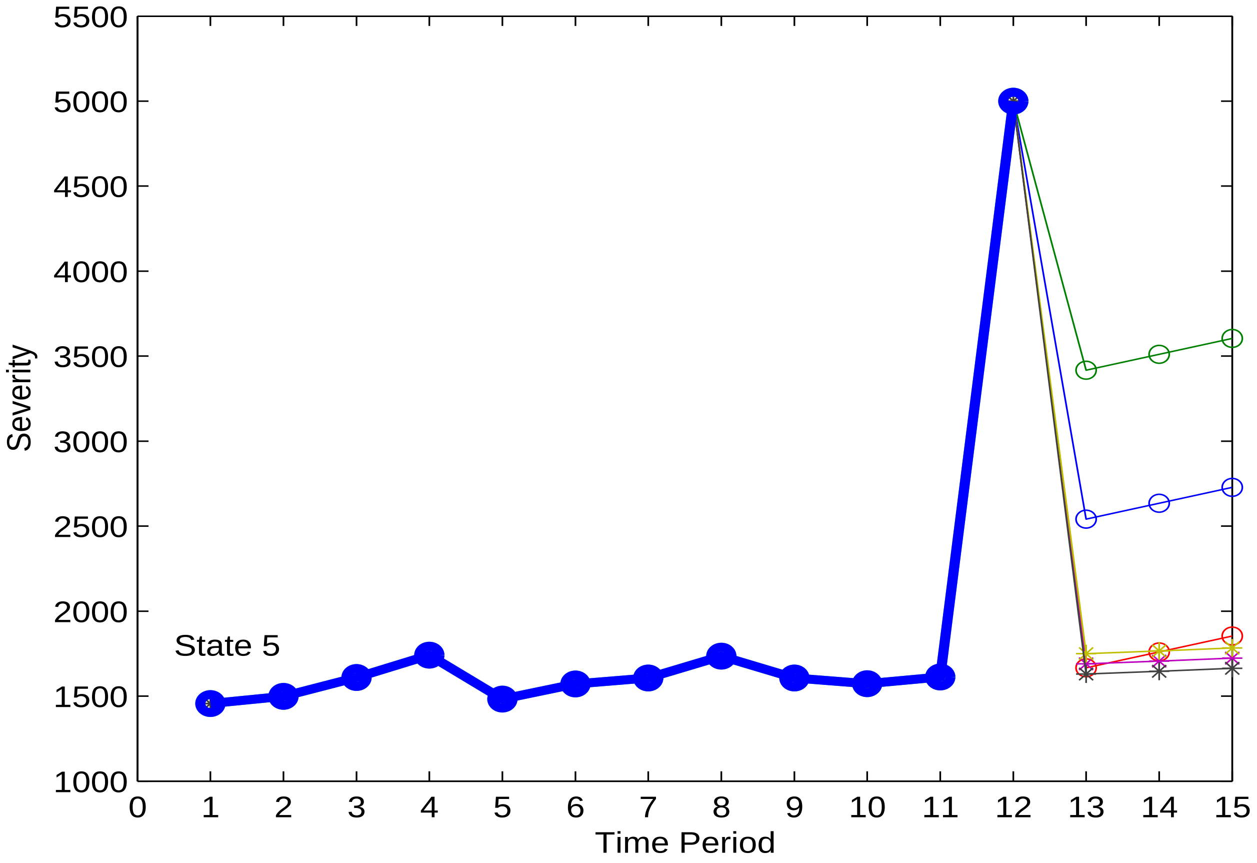

FIGURE 3: 1-, 2-, 3-step predictions for State 5 (contaminated). (CATL: marked by $*$. REML: marked by $\circ$.)

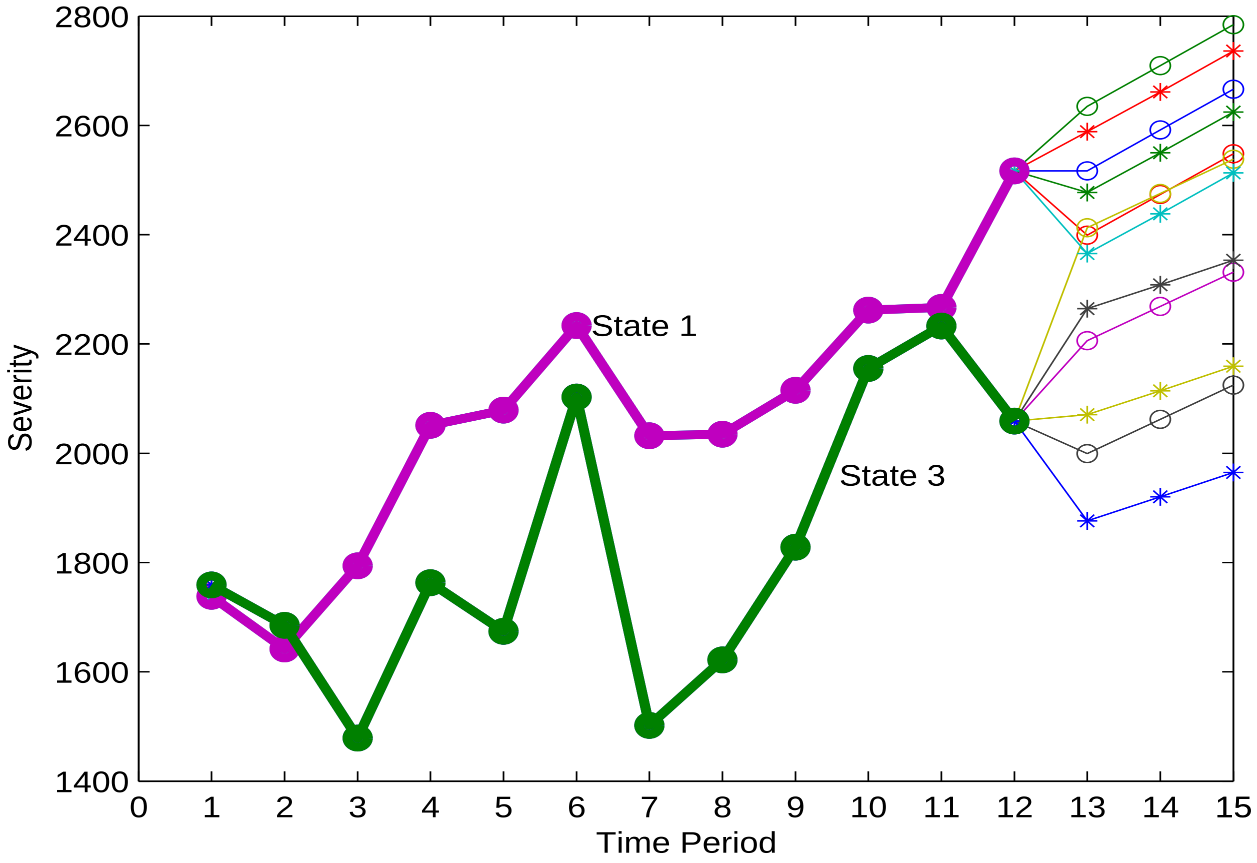

FIGURE 4: 1-, 2-, 3-step predictions for States 1, 3 (not contaminated). (CATL: marked by $*$. REML: marked by $\circ$.)

Medicare Data

- **Data**
  - $*$ 6 years (from 1990 to 1995) of claims for inpatient hospital charges; charges covered by the Medicare program in 54 states across the U.S. (including the D.C., Puerto Rico, Virgin Islands, "Other").
  - $*$ RESPONSE VARIABLE: covered claims per discharge (CCPD).
- **Model**
  - $*$ $\mathbf{x}_{it} = (1, t, \text{AVE\_DAYS}_{it})$, $\mathbf{z}_{it} = (1, t)$, where AVE_DAYS denotes average hospital stay per discharge in days (Frees, Young, Luo, 2001).
- **Objectives**
  - $*$ Model fitting; outlier identification; next-period predictions.

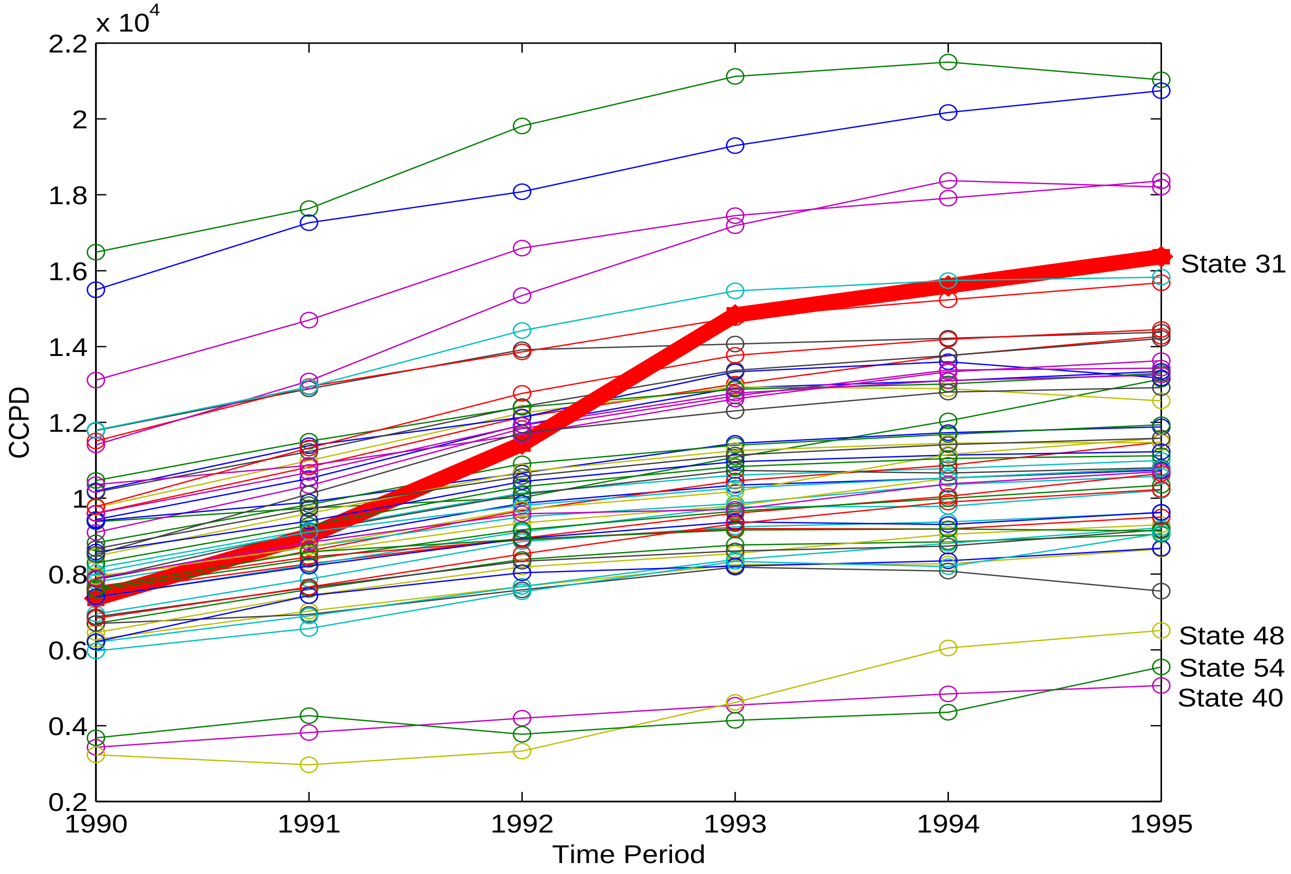

FIGURE 5: Multiple time series plot of covered claims per discharge.

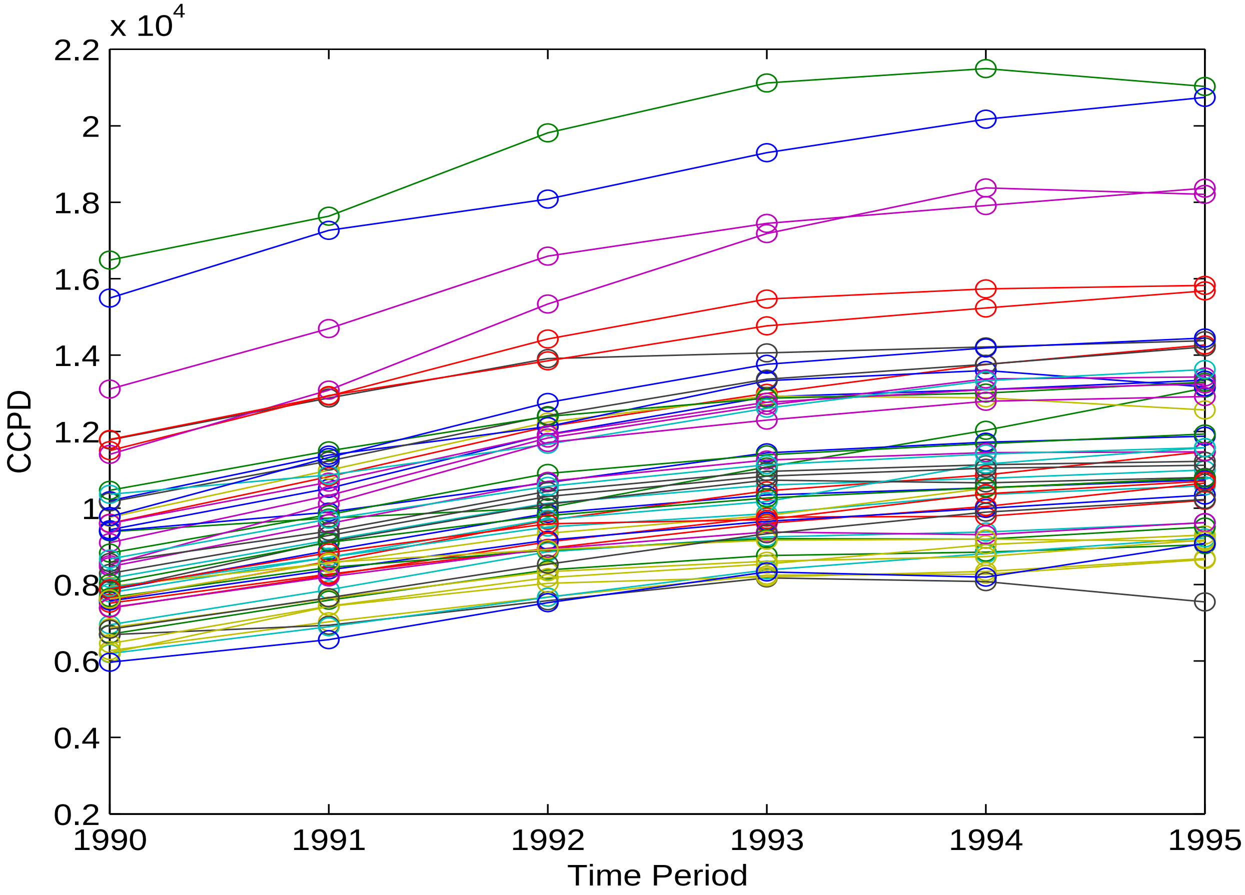

FIGURE 6: Multiple time series plot of CCPD after data cleaning.

TABLE 4: Comparison of various regression credibility models.

Model	*Intercept*	*Year*	*Year* (*State 31*)	*AVE_DAYS*
	β_1	β_2	β_4	β_3
MODEL 6	4,827	753.1	1,540.81	348.3
REML 1	7,932.6	675.3		21.9
REML 2	5,447.9	744.5		301.8
CATL 1	8.491	0.081		0.061
CATL 2	8.504	0.081		0.059

- **Model Fitting**
  - ∗ MODEL 6: Manual data cleaning; state-specific predictor for State 31, New Jersey (Frees, Young, Luo, 2001).
  - ∗ REML 1, CATL 1: Original data, no manual data cleaning.
  - ∗ REML 2, CATL 2: Manual data cleaning, but no state-specific predictor.

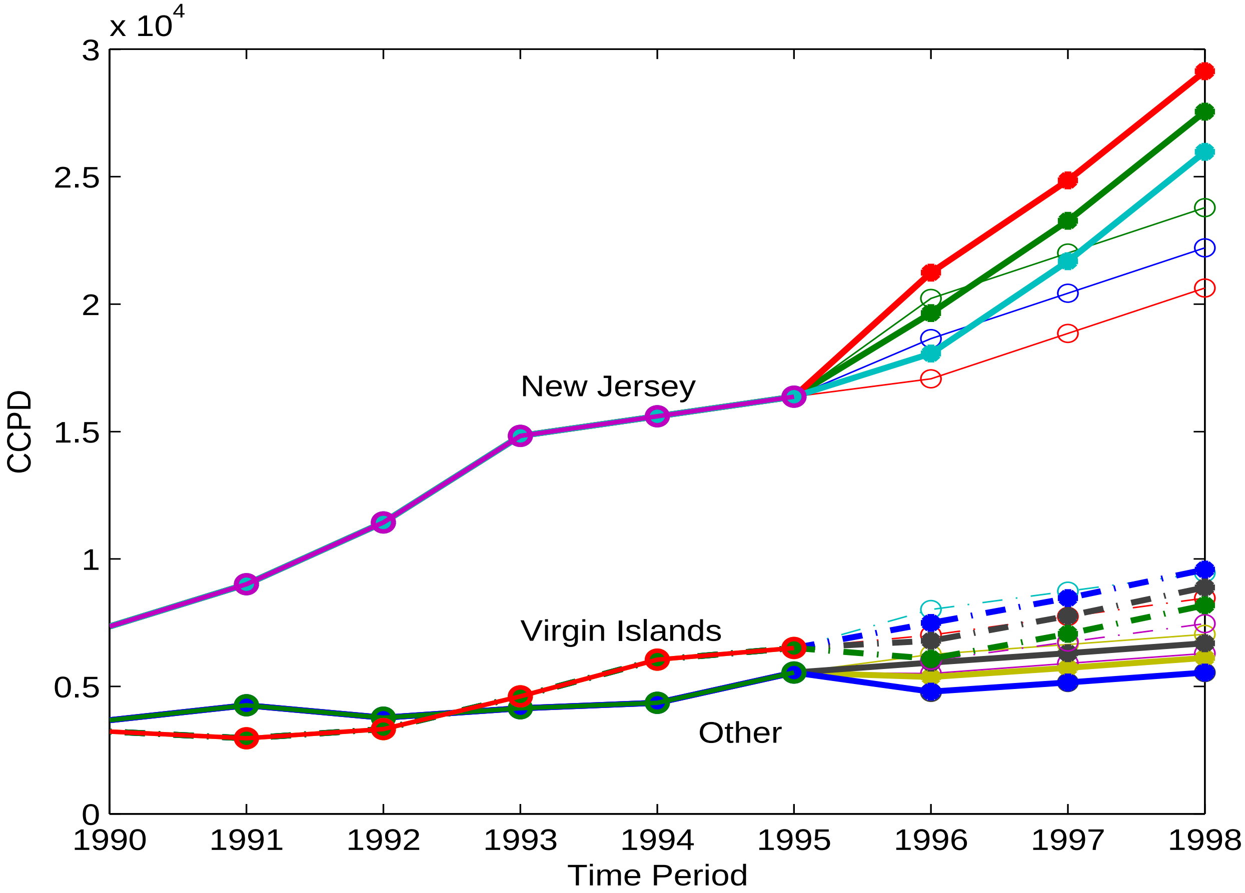

FIGURE 7: 1-, 2-, 3-step predictions for States 31, 48, 54.
(CATL: thin line, marked by ○. REML: thick line, marked by ●.)

Real Estate Data

- **Data**
  - ∗ 9 years (from 1986 to 1994) of housing sales price data reported for 36 metropolitan statistical areas (MSA) in the U.S.
  - ∗ RESPONSE VARIABLE: NARSP represents the MSA's average sale price in logarithmic units; based on transactions reported through the Multiple Listing Service, National Association of Realtors.
- **Model**
  - ∗ $\text{NARSP}_{it} = \mu + \beta_1\,\text{PERYPC}_{it} + \beta_2\,\text{PERPOP}_{it} + \beta_3\,\text{YEAR}_t + \alpha_i + \varepsilon_{it}$,

    where PERYPC denotes annual percentage growth of per capita income and PERPOP is annual percentage growth of population.
- **Objectives**
  - ∗ Model fitting; outlier identification.

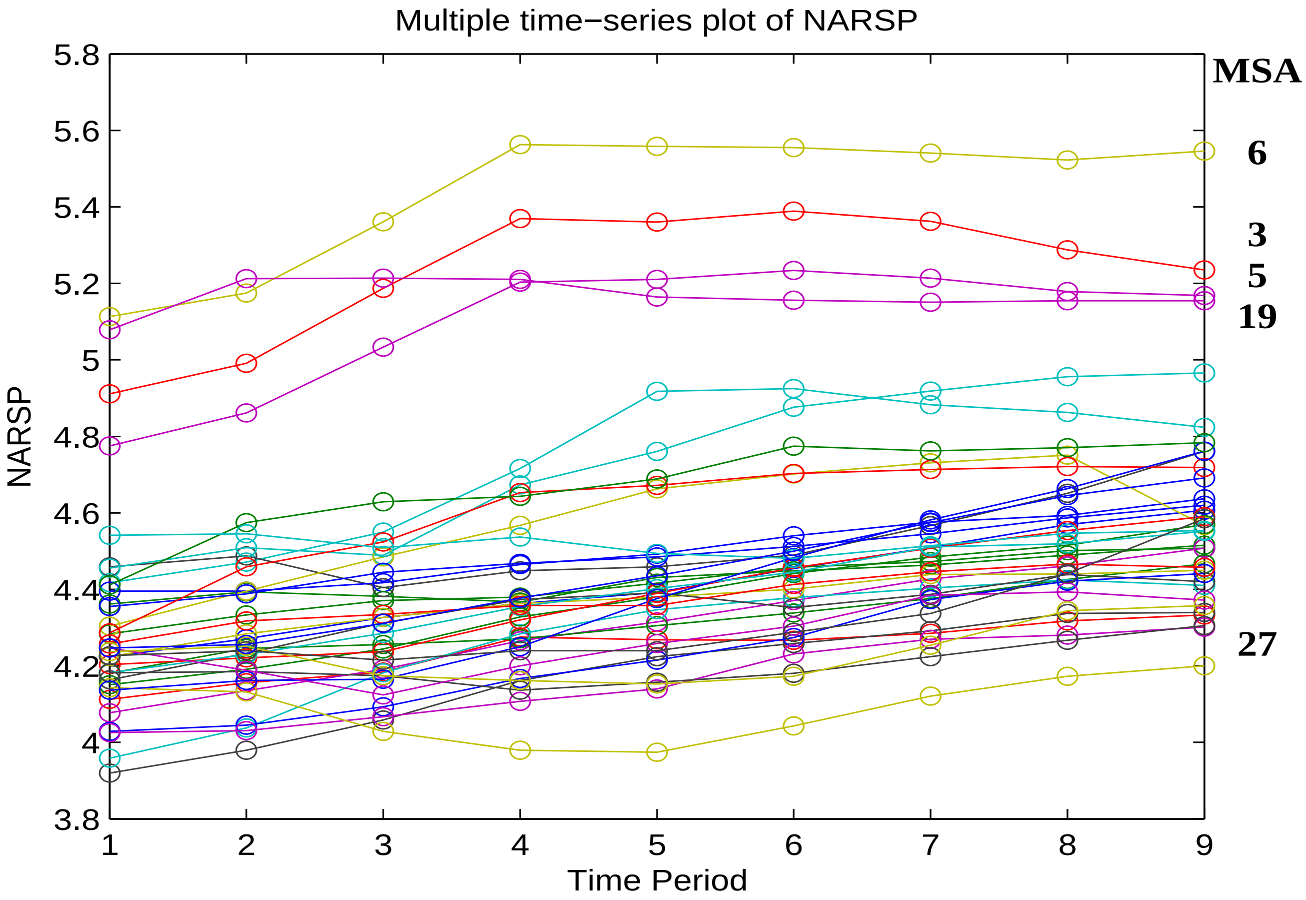

FIGURE 8: Multiple time series plot of NARSP.

TABLE 5: Estimated parameters for the housing sales price data.

Estimation	*Fixed Effects*				*Variance Components*	
Procedure	$\widehat{\mu}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_\varepsilon$	$\widehat{\sigma}^2_\alpha$
REML	4.34	-0.01	-0.00	0.04	0.005	0.097
CATL 1	4.22	-0.01	0.00	0.04	0.003	0.018
CATL 2	4.27	-0.01	-0.00	0.04	0.004	0.026

- **Model Fitting, Outlier Detection**
  - ∗ REML: Identically distributed residuals, $\varepsilon_{it} \sim N(0, \sigma^2_\varepsilon)$.
  - ∗ CATL 1: Identically distributed residuals, $\varepsilon_{it} \sim N(0, \sigma^2_\varepsilon)$.
    Outlying MSAs: #3, #4, #5, #6, #11, #19, #27, #28, #29, #32.
  - ∗ CATL 2: Non-identically distributed residuals, $\varepsilon_{it} \sim N(0, \sigma^2_{\varepsilon_i})$.
    Outlying MSAs: #3, #5, #6, #19.

4. Final Remarks

- **Summary**

  Illustrated how corrected adaptively truncated likelihoods, CATL, can be used for robust-efficient fitting of mixed linear models with fat-tailed data.

- **About CATL Methods**

  ∗ Allow to mitigate heteroscedasticity through explicit incorporation of weighting and/or use of logarithmic transformation.

  ∗ Provide robustness against outliers occurring both within and between risks through adaptive detection rules that automatically identify and reject excess claims in samples of small size.

  ∗ Yield robust credibility premiums.

  ∗ Compete well against established robust methods.