Technological Improvements or Climate Change?

– Bayesian Modeling of Time-Varying Conformance to Benford's Law.

Junho Lee* and Miguel de Carvalho

School of Mathematics, University of Edinburgh, Edinburgh, U.K. * j.lee-63@sms.ed.ac.uk

1. Introduction

Benford's Law is an empirical observation on the distribution of first digits of numerical data. The law states that the frequency of the first digit of data follows a logarithmically decreasing distribution given by

$$\mathsf{p}_d = \mathsf{P}(D = d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, \ldots, 9, \quad D: \text{ the first significant digit}$$

Benford's Law has been recently advocated as a natural tool to assess the quality and homogeneity of large datasets spanning several decades [1]. This study develops a Bayesian time-varying model that tracks dynamics of the first-digit distribution and evaluates the compliance with the Benford's Law. We apply the model to the global tropical cyclone data. Our goals are to (1) learn about the dynamics of the leading digits, (2) examine conformance to Benford's Law, and (3) assess the homogeneity within the dataset.

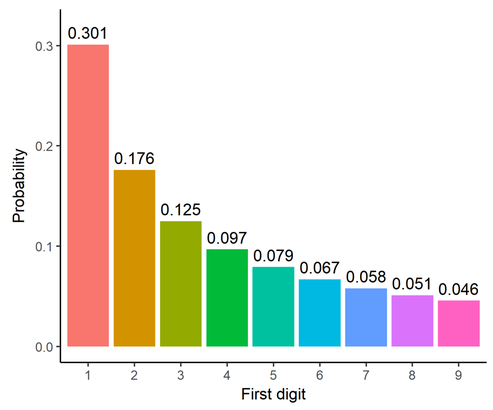

Fig 1. **Benford's Law**

2. Model

- **(Multinomial Logistic Model)** Define N_t as a total number of trials in time t and $\boldsymbol{n}_t = (n_{1,t}, \ldots, n_{9,t})$ as a random vector such that each $n_{d,t}$ denotes a frequency of event that the first digit equals to d during year t.
- **(Penalized Splines)** For inference, we follow a Bayesian version of penalized spline approach [2, 3] so as to learn about dynamics of first-digit probability.

Bayesian Smooth Multinomial Model

$$\begin{aligned}
\text{(Sampling distribution)} \quad & (n_{1,t}, \ldots, n_{9,t}) \sim \text{Mult}(N_t, p_{1,t}, \ldots, p_{9,t}), \\
\text{(Model Specification)} \quad & p_{d,t} = \frac{\exp(\eta_{d,t})}{1 + \sum_{d=1}^{8} \exp(\eta_{d,t})}, \quad p_{9,t} = \frac{1}{1 + \sum_{d=1}^{8} \exp(\eta_{d,t})}, \\
& \eta_{d,t} = \sum_{k=1}^{K+3} \beta_{d,k} B_{d,k}(t), \\
\text{(Random Walk Prior)} \quad & \beta_{1,d} \sim U(c_0, d_0), \quad \beta_{k+1,d} = \beta_{k,d} + \varepsilon_d, \quad \varepsilon_d \sim N(0, \tau_d^2), \\
\text{(Hyper-Prior)} \quad & \tau_d^2 \sim \text{IG}(a_0, b_0).
\end{aligned}$$

3. Comparison

The Global Tropical Cyclone (GTC) dataset includes a register of tropical cyclones around the world since 1842. The dataset

- provides information on a geographical location, frequency, and intensity of each cyclone.
- stimulates a debate on the quality of early records for assessing the influence of climate change on the occurrence of cyclones [4].

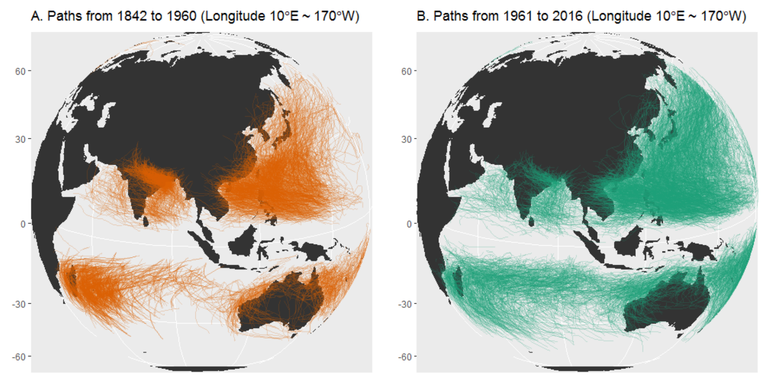

Fig 2. **Map of tropical cyclone tracks since 1842**

We calculate a traveled distance of each cyclone and analyze the distribution of the leading digits of 12,741 cyclones until 2016.

Th first-digit distribution of the pooled data resemble the probabilities from Benford's Law.

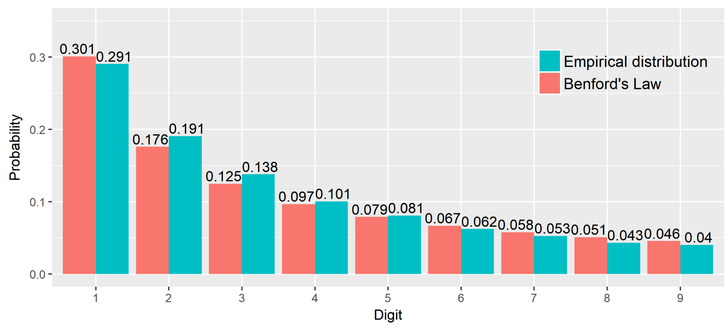

Fig 3. **The first-digit distribution of the pooled data**

4. Data application

The posterior of the first-digit probability presents different patterns among digits over the period.

- The proportions of digit one/three shows a profound deviation from Benford's Law with a persistently decreasing/increasing trend.
- The other curves move together tightly around the probabilities from the First-digit Rule.
- Reflecting small sample sizes, the credible bands of the early stage are much wider than those in the period of 1900s onward

We use a smooth Sum of Squared Deviation (SSD) to examine homogeneity within the dataset over time. $\text{SSD}(t) = \sum_{d=1}^{9} (p_{d,t} - \mathsf{p}_d)^2$

- Our method avoids a discretization effect from empirical distribution approaches, which misleads to lack of conformance to Benford's Law if the sample size is small.

Our analysis suggests:

- A period heterogeneity exists from 1880 to 1940, possibly due to incomplete management of cyclone records and inevitable measurement errors.
- The technological improvements may have had a moderate influence on the homogeneity of the dataset, and recent heterogeneity could be due to other drivers such as climate change.

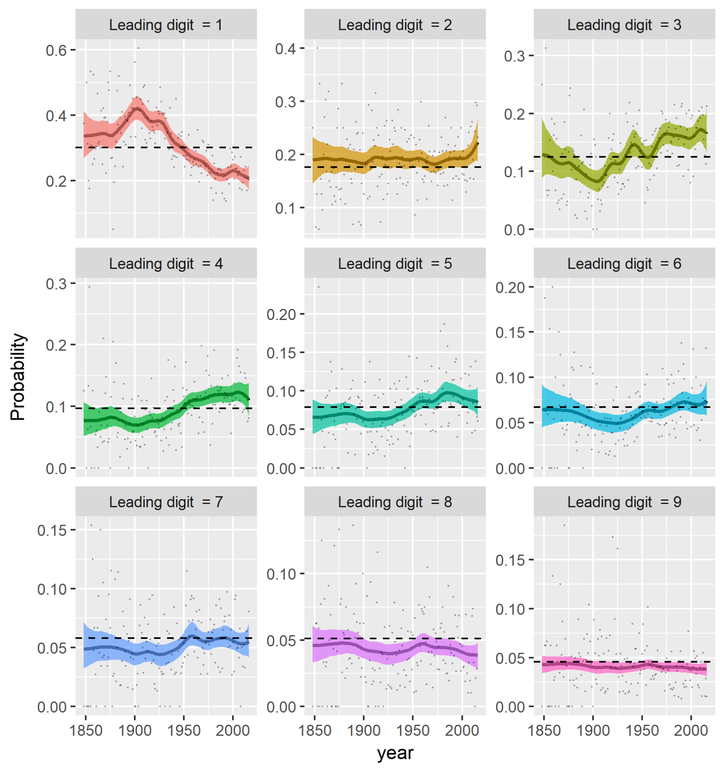

Fig 4. **Dynamics of** $(p_{1,t}, \ldots, p_{9,t})$

The posterior mean (solid line) and 95% credible bands (shaded area), the empirical distribution (point), and Benford distribution (dashed line).

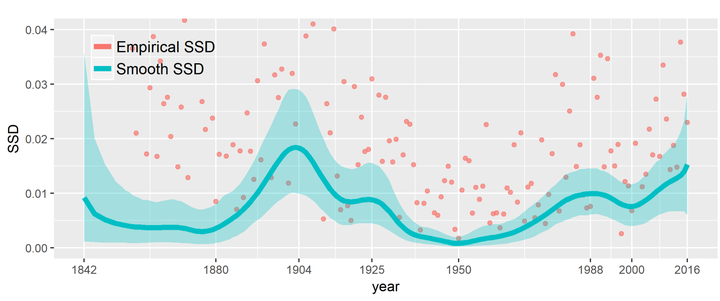

Fig 5. **Homogeneity of GTC data**

The posterior mean of SSD (solid blue line), 95% credible bands (shaded blue area) and the empirical SSD (red point) in each year.

5. References

[1] Steven J. Miller. *Benford's Law: Theory & Applications*. Princeton University Press, 2015.

[2] Stefan Lang and Andreas Brezger. Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

[3] Andreas Brezger and Winfried J. Steiner. Monotonic regression based on Bayesian p-splines: An application to estimating price response functions from store-level scanner data. *Journal of Business and Economic Statistics*, 26:90–104, 2008.

[4] Renaud Joannes-Boyau, Thomas Bodin, Anja Scheffers, Malcolm Sambridge, and Simon Matthias May. Using Benford's law to investigate natural hazard dataset homogeneity. *Scientific Reports*, 5, Jul 2015.

[5] Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical Science*, 10(4):354–363, 1995.