Spatio-Causal Patterns of Sample Growth

ANDRE F. RIBEIRO*, Harvard University, USA and University of Sao Paulo, Brazil

Different statistical samples (e.g., from different locations) offer populations and learning systems observations with distinct statistical properties. Samples under (1) 'Unconfounded' growth preserve systems' ability to determine their variables' effects on outcomes-of-interest (and lead, therefore, to interpretable black-box predictions). Samples under (2) 'Externally-Valid' growth preserve their ability to make predictions that generalize across out-of-sample variation. The first generates predictions that generalize over sample populations, the second over their common unobserved factors. We illustrate these theoretic patterns in the full American census from 1840 to 1940, and samples ranging from the street-level all the way to the national. This reveals new conditions for the generalizability of samples over space and time, and connections among the Shapley value, counterfactual statistics, and hyperbolic geometry.

JAIR Associate Editor: Quanquan Gu

JAIR Reference Format:
Andre F. Ribeiro. 2025. Spatio-Causal Patterns of Sample Growth. *Journal of Artificial Intelligence Research* 83, Article 19 (July 2025), 7 pages. DOI: 10.1613/jair.1.15675

1 Appendices

2 Methods

The only sample selection criteria employed, across years, was 'all males in working age' (between 14 and 55, not-included, and not in school). Locations are lat-lon coordinates of sample units' enumeration district. Enumeration districts are the census finest spatial resolution, above only individuals' free-form street address. Fig.3(b) illustrates districts for NYC in 1880.

For Fig.3(c-h), we assemble samples at increasing spatial levels, for each enumeration district in the US (a 'location'). For a fixed location x_0, the sample for level s_t consists of all districts (thus all their sample units) within distance s_t from x_0. For level t this distance is $t \times \Delta s$, for a constant Δs. These are determined from the spatial extension (state or national) and number of levels in each task (mentioned in the main text). Occupation frequency at a given level s_t are counts over the corresponding sample. For the time-series and spectral results, frequency vs. level plots were assembled for increasing Δs from the district to the national level, Fig.3(g,h). Autocorrelation is the correlation of these scaling spatial-series, for each individual location and occupation. This is the correlation between the present level s_t and the previous level s_{t-1}, then with all antecedent s_{t-2}, s_{t-3}, etc. Namely, the auto-correlation function (ACF) of the spatial-level series is

$$\rho_k = \frac{\sum (n_{s_t} - \mu) \times (n_{s_{t-k}} - \mu)}{\sigma_{n_{s_t}} \times \sigma_{n_{s_{t-k}}}} = \frac{\mathbf{E}\Big[(n_{s_t} - \mu) \times (n_{s_{t-k}} - \mu)\Big]}{\sigma_{n_{s_t}} \times \sigma_{n_{s_{t-k}}}}, \tag{S1}$$

*Corresponding Author.

Author's Contact Information: Andre F. Ribeiro, ribeiro@alum.mit.edu, Harvard University, Cambridge, Massachusetts, USA and University of Sao Paulo, Sao Carlos, Sao Paulo, Brazil.

DOI: 10.1613/jair.1.15675

for a lag of k (x-axis), and mean and standard deviations μ and σ. For a set of locations x_0, a factor a's min. and max-frequency rank across locations is taken as $r_0 = \min[r(a, x_0)]$ and $r_\omega = \max[r(a, x_0)]$, where $r(a, x_0)$ is the rank of factor a in location x_0.

Fig.S1(b) shows all states with non-hyperbolic auto-correlation in 1880. Fig.S2 reproduces Fig.2(h), auto-correlation vs. spatial level, for states not in the main article (all years and occupations). Fig.S3 reproduces Fig.3(g), min. and max. frequency ranks across levels.

Fitting of univariate distributions to non-censored data in Fig.4(e) is done by moment matching (mme) [2, 1]. Theoretical and empirical moments are matched by minimizing the sum of squared differences between observed and theoretical moments. The Pareto distribution has two parameters, shape a and scale s, and density $y(x) = as^a/(x+s)^{(a+1)}$ for $x > 0$, $a > 0$ and $s > 0$. The coth method constitutes of fitting directly a coth function with angle and multiplier as parameters, $y(x) = a\coth(mx)$. Fig.4(e) shows ratio in goodness-of-fit between the Pareto and coth models, for all years and spatial levels, according to a Bayesian information criterion (BIC). Absolute BIC values are generally not informative, but their ratios are popular means of comparing the fit offered by alternative models. Models with lower BIC are preferred. Since the two models have the same number of parameters, the ratios indicate relative fitness of the two previous functional forms. It illustrates the gains derived from modeling both min. and max. ranks in heterogeneous locations.

The Pareto distribution is often used to describe quantities over a fixed threshold. It is therefore appropriate to describe the case of overcomplete squares (i.e., above s_{sq}). Let $P([\frac{\omega}{n}e]^m > \omega) = 1 - (\frac{\omega}{n}e)^m$. If n is exponentially distributed with rate m, then ze^n is Pareto with minimum z and shape m. Regressing the shape, m, and minimum, $z = \frac{\omega}{n}^m$, parameters of this Pareto distribution allow us to estimate $\frac{\omega}{n} = z^{1/m}$, and constitutes a second method to demonstrate the limits discussed in the main article. The estimated value of ω/n for New York, Philadelphia, Iowa and Maryland, Fig.3(h), are 0.81 ± 0.091, 0.81 ± 0.012, 0.81 ± 0.037, and 0.82 ± 0.154 (mean and standard deviation across years). Estimates for other states are indicated in Fig.S2 (above boxes). These predicted values for ω/n across spatial-levels, using other techniques, agree with the previous results from time-series autocorrelation, hyperbolic regression and spectral methods. The coth plots in Fig.4(h) are generated by scaling max-rank and min-rank values by, respectively, m/e and $m \times (1 - /e)$, and summing m/e to the latter.

We illustrated the formulated patterns in Economic-related variables available in the census with catenaries (such as the distribution of occupations, and the highly correlated, distribution of industries, across locations). Although we focused there on Economic variables, we observe the same patterns in subsets of demographic variables with obvious effects on economic growth, according to the Economic literature. Fig.S1(a) shows the example of the 10 variables with highest effect in the classification task in Fig.4(g). They were selected with a Shapley-value importance estimate [3, 5] of the algorithm with highest average accuracy across all states and years, using the 10.055 variables census dataset. The 10 demographic-only variables with highest effect according to this procedure were: age, birthplace, metropolitan location, marriage status, literacy status, african-american race, white race, hispanic race, size of place, family size. See [4] for examples with smaller-scale demographic census data, but across other domains.

3 Hyperbolic Geometry and Permutations

The main article presented a connection between the Combinatorics quantities used and hyperbolic geometry, based both on Pascal's triangle and a generic sample growth process. An alternative connection can be reached from Taylor's expansion of trigonometric functions, as noted in Eq.(6-7). Eq.(6)(main text) starts with a known relationship among permutations, combinations and derangements. From their respective definitions,

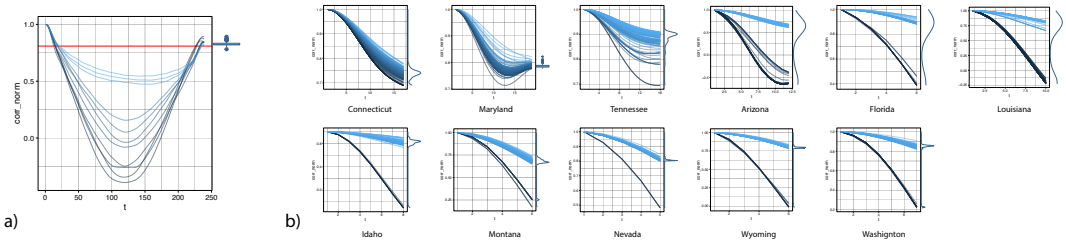

Fig. S1. (a) Auto-correlation vs spatial level for 10 demographic (non-economic) factors with large effects (Pensylvia, 1880), (b) further example states with subexponential auto-correlation, in 1880.

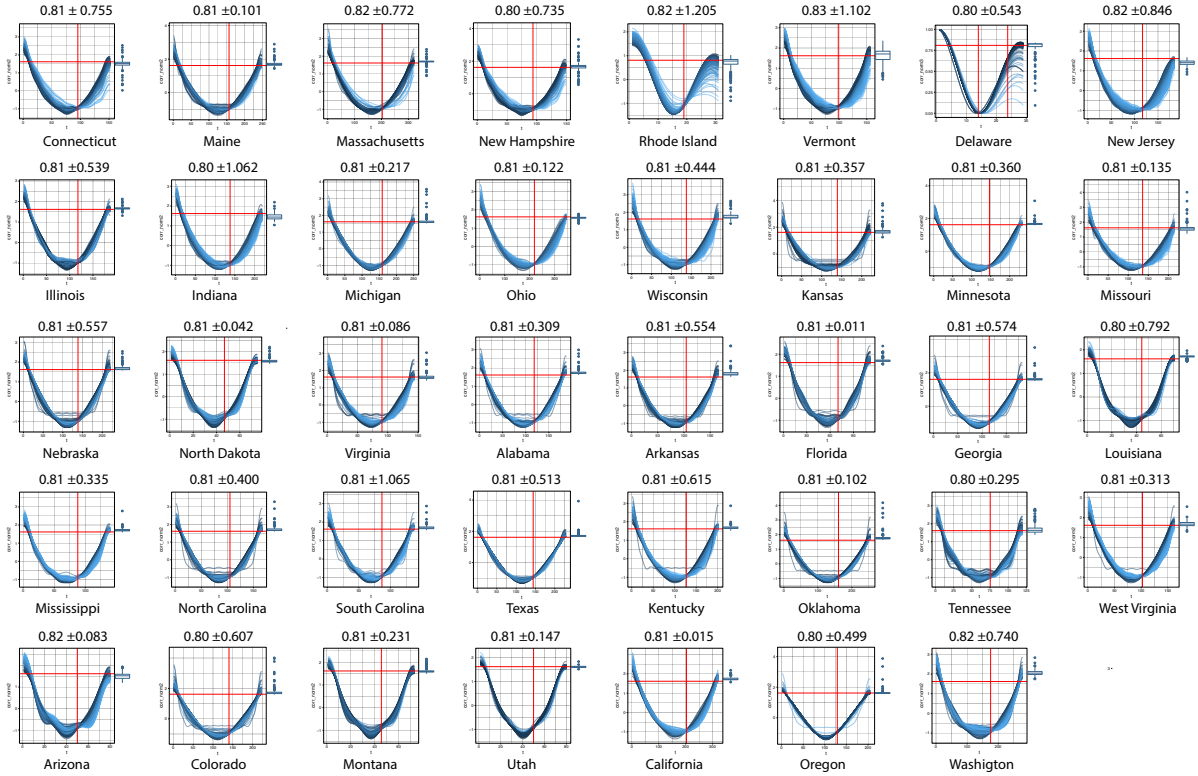

Fig. S2. Auto-correlation of spatial levels (states not in the main article), across all years and locations, boxplot of catenary slag (sidepanel), 0.809 correlation and $m \times (1 - 1/e)$ indicated (red line), Pareto-based alternative estimate, and their standard deviation, across year and states (above boxes) .

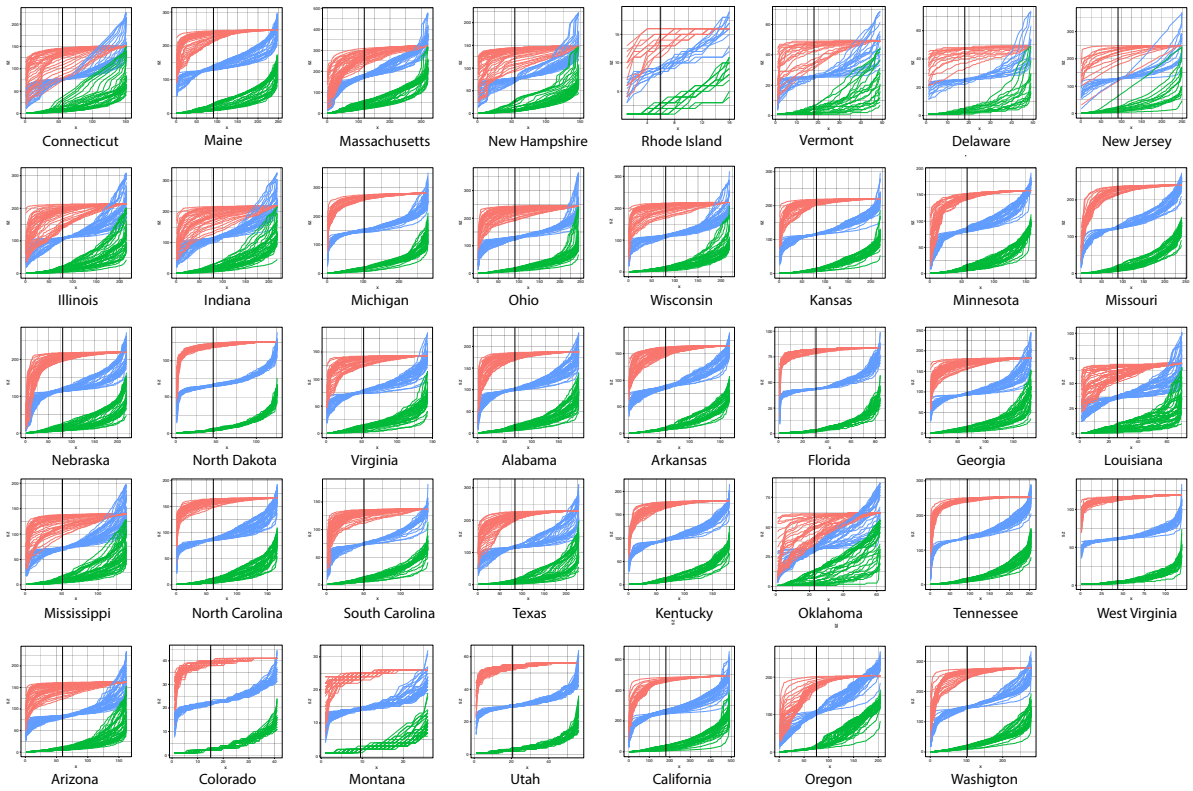

Fig. S3. Frequency rank of all spatial levels (states not in the main article).

$$
\begin{aligned}
m! &= \sum_{t=0}^{m} \left[\binom{m}{t} \times D_{m-t} \right], \\
&= \sum_{t=0}^{m} \left[\frac{m!}{(m-t)!t!} \times (m-t)! \sum_{i=0}^{m-t} \frac{(-1)^i}{i!} \right], \\
&= \sum_{t=0}^{m} \left[\frac{m!}{t!} \times \sum_{i=0}^{m-t} \frac{(-1)^i}{i!} \right].
\end{aligned}
$$

A falling factorial is $(m)_j = m \times (m-1) \times ... \times (m-(j-1))$. Derangements are the permutations, of a size m, with highest amount of variation (with no fixed-points). Falling factorials offer an alternative definition for partial permutations, sometimes used, where overlaps are allowed in the partial permutation's varying section (instead of derangements, as in squares, and most usually). Introducing falling factorials in the previous equation,

$$
\begin{aligned}
m! &= \sum_{t=0}^{m} \left[(m)_{m-t} \times \sum_{i=0}^{m-t} \frac{(-1)^i}{i!} \right], \\
&= \sum_{t=0}^{m} \left[\sum_{i=0}^{m-t} \frac{(-1)^i \times (m)_{m-t}}{i!} \right], \\
&= \sum_{t=0}^{m} \left[\sum_{i=0}^{m-t} \frac{(-1)^i \times \left(\sum_{j=0}^{m-t} s(m-t,j) m^j \right)}{i!} \right], \qquad \text{(S2)} \\
&= \sum_{t=0}^{m} \left[\sum_{j=0}^{m-t} \sum_{i=0}^{m-t} s(m-t,j) \frac{(-1)^i m^j}{i!} \right]. \qquad \text{(S3)}
\end{aligned}
$$

Eq.(S2) follows from Stirling's definition for falling factorials. Its coefficients, $s(m,j)$, are the Stirling numbers of the first kind (the number of cycles of size j, in permutations of size m). The two inner sums (index i and j) in Eq.(S3) create $(m-t)^2$ terms in each outer iteration (indice t). Organizing these terms in a square matrix of size $(m-t)$,

$$
\begin{array}{c|ccccc}
i/j & 1 & 2 & 3 & \\
1 & \frac{(-1)^0 m^0}{0!} s(m-t,0) & \frac{(-1)^0 m^1}{0!} s(m-t,1) & \frac{(-1)^0 m^2}{0!} s(m-t,2) & \cdots \\
2 & \frac{(-1)^1 m^0}{1!} s(m-t,0) & \frac{(-1)^1 m^1}{1!} s(m-t,1) & \frac{(-1)^1 m^2}{1!} s(m-t,2) & \cdots \\
3 & \frac{(-1)^2 m^0}{2!} s(m-t,0) & \frac{(-1)^2 m^1}{2!} s(m-t,1) & \frac{(-1)^2 m^2}{2!} s(m-t,2) & \cdots \\
 & \vdots & \vdots & \vdots &
\end{array},
$$

shows that diagonals contain the exact definitions of hypergeometric functions, as Taylor series (two diagonals are illustrated, in red and blue, above), then,

$$
\begin{aligned}
m! &= \Big[\cosh(m) - \sinh(m)\Big] \times \sum_{t=0}^{m} \left[(m{-}t) \times \left(\sum_{k=0}^{m-t} s(m{-}t,k)\right)\right], \\
&= \Big[\cosh(m) - \sinh(m)\Big] \times \sum_{t=0}^{m} \Big[(m{-}t) \times (m{-}t)!\Big], \\
&= \Big[\cosh(m) - \sinh(m)\Big] \times [(m{+}1)! - 1], \\
&= \frac{(m{+}1)! - 1}{\cosh(m) + \sinh(m)},
\end{aligned}
\tag{S4}
$$

The factor not included in hyperbolic definitions (but common across diagonals), Signed Stirling numbers, sum to the factorial of their upper limit, Eq.(S4). Finally, using $\sum_{t=0}^{m} t \times t! = (m+1)! - 1$, De Moivre's theorem and trigonometric sign symmetries, leads to Eq.(7) (main text). Notice that because the two hyperbolic functions share a constant angle, m, it is implied that $tanh(m)$ must be constant for hyperbolic trigonometric functions to count all permutations.

4 Permutation Enumeration rate (ω) and Hyperbole

The following relationship is known but reviewed here for convenience. The equation $x \times y = \omega$, represents a hyperbola when expressed in standard form. A hyperbola is a conic section with two branches, and its standard equation typically has the form,

$$\left(\frac{x^2}{a^2}\right) - \left(\frac{y^2}{b^2}\right) = 1. \tag{S5}$$

In this equation, a and b are positive constants that determine the shape and orientation of the hyperbola. To demonstrate the connection, first divide both sides of $x \times y = \omega$ by ω,

$$\left(\frac{x \times y}{\omega}\right) = 1. \tag{S6}$$

Express this in the form of a product of squares,

$$\left(\frac{x}{\sqrt{\omega}}\right) \times \left(\frac{y}{\sqrt{\omega}}\right) = 1. \tag{S7}$$

Introduce new constants a and b such that $a = \sqrt{\omega}$, $b = \sqrt{\omega}$ and $x/a \times x/b = 1$, which leads to the additive relationship,

$$\left(\frac{x^2}{a^2}\right) - \left(\frac{y^2}{b^2}\right) = 1. \tag{S8}$$

The constants a and b (and thus ω) will determine the specific properties of the hyperbola. The parametric equations for its right branch are $x = a \times \cosh(t)$ and $y = b \times \sinh(t)$.

References

[1] Marie Laure Delignette-Muller and Christophe Dutang. "fitdistrplus : An R Package for Fitting Distributions". In: *Journal of statistical software* 64.4 (2015), pp. 1–34. DOI: 10.18637/jss.v064.i04.

[2] H. Christopher Frey and Alison C Cullen. *Probabilistic techniques in exposure assessment : a handbook for dealing with variability and uncertainty in models and inputs*. New York: Plenum Press, 1999. ISBN: 0306459566; 0306459574.

[3] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[4] Andre F Ribeiro. "Competition, Diversity and Quality". In: *Physica A* 568 (2021), p. 125683. DOI: 10.1016/j.physa.2020.125683.

[5] Andre F. Ribeiro. "Sample observed effects: enumeration, randomization and generalization". In: *Scientific Reports* 15.1 (2025), p. 8423. DOI: 10.1038/s41598-024-80839-8. URL: https://doi.org/10.1038/s41598-024-80839-8.

Received 28 February 2023; revised 18 July 2023; accepted 26 July 2024