

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 22 (July 2025), 19 pages. doi: [10.1613/jair.1.15675](https://doi.org/10.1613/jair.1.15675)

1 Introduction

Large scale and high-dimensional geospatial datasets currently offer rich opportunities for predictive and Geo-AI applications [32, 15, 19, 33] (e.g., disease incidence, ecological behavior, electoral results, crime occurrence, economic growth, recommendation systems). While it is common practice to train regression and classification models in data collected across distinct locations, little is known about how their out-of-sample accuracy ('predictiveness') and biasedness (e.g., black-box 'fairness') [7, 34] are expected to change across spatial extensions. The first indicates whether predictions derived from the sample will be close to their true values for a population in conditions different from those at data collection (i.e., whether they will 'generalize'), and the latter whether they will systematically favor individual populations (e.g., as result of insufficient sample sizes, unobserved variables, or other failures in sample selection). Understanding these issues is important because they allow us to answer crucial questions for learning systems: Can predictions made for a given population, with data from one location, be used in others? Does collecting larger samples improve prediction accuracy for that first population? We first formulate functions describing fairness-generalizability tradeoffs across space, revealing their connections to hyperbolic geometry and theoretic experimental designs. We then consider 100 years of the American census (and all variables in the census) as case study. For each cross-section (decade), we consider the important task of predicting economic growth for over 60K individual locations under increasing spatial samples. We demonstrate how (1) generalizability tradeoffs evolve across spatial levels, and (2) repeat the validation of generalizability limits derived in [27] for the spatial domain, and with the current census micro-data.

Let $S : X_m \rightarrow [0, 1]$ describe any learning system or agent using an input sample X_m with m variables to derive a classification decision, $S(X)$. Our central goal is to formulate how the generalizability of these systems changes

*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.



This work is licensed under a [Creative Commons Attribution International 4.0 License](https://creativecommons.org/licenses/by/4.0/).

© 2025 Copyright held by the owner/author(s).

doi: [10.1613/jair.1.15675](https://doi.org/10.1613/jair.1.15675)

across space (i.e., to what extent a model assembled in a location will hold for others), and, thus, to identify a parametric functional form \mathcal{F} that can describe accuracy bounds across possible S ,

$$\mathcal{F}(d_{x_0}) = \max_S \left\{ \text{ACC}(S; X[x_0, d_{x_0}]) \right\}, \quad (1)$$

where ACC indicates the accuracy of models trained in samples $X[x_0; d_{x_0}]$ encompassing all observations at distances less than or equal to d_{x_0} from x_0 . The **specific way in which \mathcal{F} changes across space offer limits and opportunities to algorithmic and agent learning systems and their performance**. Strict bounds on the uncertainty of predictions afforded to algorithms or agents using local data can have a profound impact on the usefulness, scope, and quality of their strategic decisions or the recommendations they offer. A non-decreasing \mathcal{F} , with increasing d_{x_0} , indicates that S is able to learn a model that is accurate across locations, while a decreasing function indicates that learning fails to generalize across locations. A complementary issue would be to what extent S would be able to identify the effect (or importance) of individual variables to prediction using the same sample.

To illustrate these two fundamental sample characteristics, consider a set of binary attributes $X_4 = \{a, b, c, d\}$ observed across US locations, such as, respectively, recorded presence of crime, police stations, banks and ice-cream shops. The behavior of these entities are possibly interconnected, which implies that any calculated statistic $y \in \mathbb{R}$ over a is in fact an statistic over $y(a | b, c, d)$. Because of this, we say that a local sample has low External Validity (EV), since any changes in factors $\{b, c, d\}$ (or the many other factors that can conceivably affect crime), can invalidate the statistic. At the same time, because banks often appear together with ice-cream shops in commercial and affluent neighborhoods, we are not sure whether their presence plays any essential role when predicting crime incidence (i.e., whether they have only a spurious relationship to crime). Because of this, we say that such samples are confounded (CF). How can these two statistical issues be addressed and quantified? Comparison of crime incidence between this location and a second with banks but no ice creams shops, everything else constant, would lend evidence to the fact that ice cream shops are not driving crime up. These types of ideal what-if statements, where the effect of an outcome is observed under a single difference (while holding other factors constant), are often called counterfactual statements [23, 29]. In the hypothetical case where all such conditions can be observed, the problem of determining whether a factor is relevant to prediction is fairly easy to solve. These conditions have been formulated mathematically in the study of experimental designs [22]. The more challenging, and practical, aspect of this problem is, however, the case of unobserved conditions: often the relevant factors that change across locations are neither observed nor held constant. The issue of unobserved confounding is particularly serious for effect estimation, as effect statistics calculated from the sample might then also reflect the effect of 'exogeneous' or unobserved variation that cannot be easily discounted by typical regression and effect estimation methods [31, 23]. Although in these conditions effect observations are flawed or noisy counterfactual observations, we still refer to them as 'counterfactual' observations for short. We address these problems by studying spatial sample growth: we start with a sample with a single unit x_0 (all conditions unobserved), we then progressively add other units to the sample at larger distances to x_0 , d_{x_0} , progressively decreasing its number of unobserved conditions. We consider $\text{ACC}(S; X[x_0, d_{x_0}])$ across these samples, and, in particular, how EV and CF change across scales.

The two previous problems reflect two key, but distinct, learning problems [18]: supervised out-of-sample prediction and factor effect (or importance) estimation. Supervised prediction focuses on making accurate predictions of an outcome in unseen data using a training sample, while factor effect estimation aims to measure the relevance of specific factors for prediction and model selection. These problems are, however, closely connected [18, 7], as complete and correctly specified models lead to accurate predictions. We will demonstrate that combinatorial properties of samples impact these two traditional problems differently and

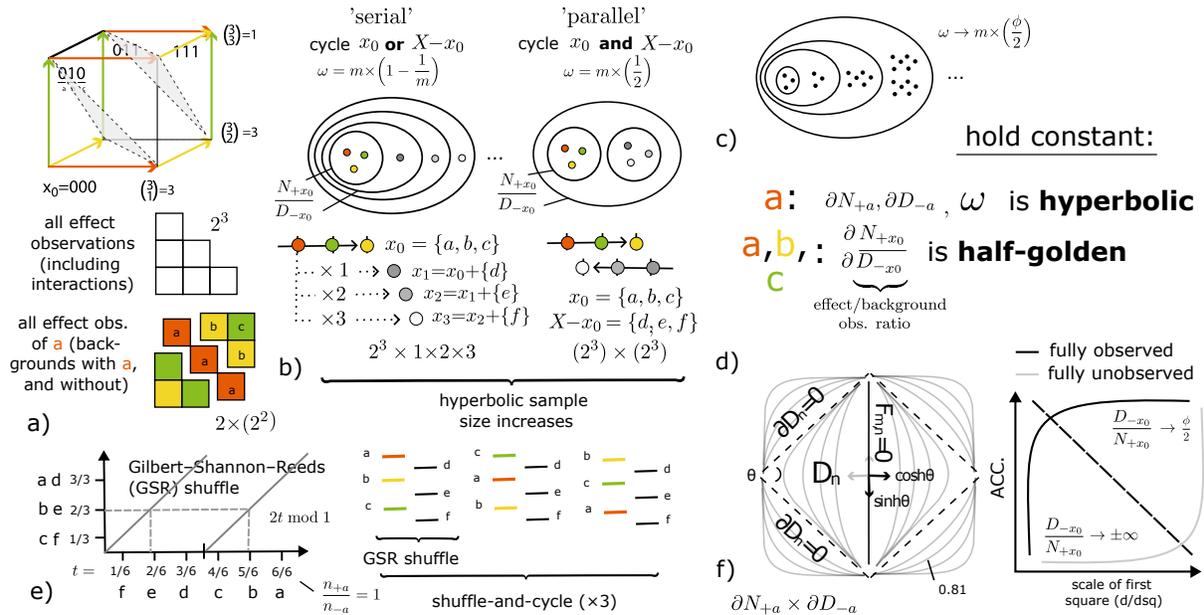


Fig. 1. Can counterfactual effect observations made in one location be used in another (are they externally valid, EV)? Can their independent effects or relevance to prediction be distinguished from others (are they unconfounded, CF)? (a) a m -dimensional hypercube and Pascal triangle with m rows portrays the full set of counterfactual effect observations with m factors in a sample X_m ($m = 3$), a $m \times m$ Latin-Square ('square') portrays all effect observation backgrounds, more counterfactual effect observations increase guarantees over the generalizability and bias of samples' effect observations, (b) 'serial' and 'parallel' interleaving of in and out of sample factors during sample growth and their expected sample sizes and growth rates, (c-d) sample sizes under EV-CF growth follow hyperbolic forms, with a half-golden rate of effect-to-background rate for the high-dimensional case, (e) illustration of Gilbert-Shannon-Reeds (GSR) shuffling as a mechanism to increase generalizability (EV) in partially-observed samples, and a shuffle-and-cycle strategy as an alternative which also guarantees decreases in factor effect confoundness (CF), (f) hyperbola for samples with distinct N_{-a}/N_{+a} (left) and expected accuracy for samples with small or large numbers of unobserved factors (right).

reveal tradeoffs across samples and locations. Fig.1(d) outlines the two central contributions of the proposed framework, which establish hyperbolic forms for $\mathcal{F}(d_{x_0})$ and consequent asymptotic limits to black-box accuracy. We first briefly formulate these contributions. After this summary, we relate the approach to known black-box importance estimation and causal effect estimators, and, finally, formulate the proposed combinatorial-geometric relationships in detail.

2 Sample Counterfactual Observation Limits

Let $X = \{a, b, c, \dots, [m]\}$ be a set of (observed or unobserved) binary factors¹ characterizing a population x , $x \in [0, 1]^m$, and $y(x)$ be a measurement over the population, $y(x) \in \mathbb{R}$. Behind many effect or importance estimation procedures is an experimental procedure (*Sect. Related Work*): add a to every variation of other populations; with each insertion, observe before-after outcome differences, $\Delta y(a)$. Consider a sample 'growth' process where we start with a fixed sample unit x_0 and, as we observe each new unit, we also observe their differences from x_0 , in respect to both x and y . Starting with a population x_0 , each possible sample growth trajectory (e.g., x_0, x_1, x_2, \dots) is an temporally or spatially ordered observation of the impact of gaining, or

¹where $[m]$ is the m -th factor in a sample with variables X_m .

losing, a set of factors on y . In other words, each step is a counterfactual effect observation for sample unit x_0 , $\Delta y(x_0 - x_1) = y(x_0) - y(x_1)$ (where the former difference is over sets and the latter over scalars). Each effect observation is thus defined by: (1) a difference in factors (what changed, $x_0 - x_1 \subseteq X_m$), an intersection of factors (what remained the same, $x_0 \cap x_1 \subseteq X_m$), and finally, a difference in outcomes (an observed scalar 'effect', $\Delta y(x_0 - x_1) \in \mathbb{R}$). For each trajectory and time, we can consider the biasedness and generalizability afforded by the accumulated samples to, for example, black-box predictors of y , and whether their performance is related to the increasing set of accumulated counterfactuals.

2.1 Sampling Effect-to-Background Ratios (ω)

The growth 'space' for any collected sample is then the imaginary space that contains all of the conceivable ways in which we could have assembled X from any one of its individual sample units x_0 (*Sect. Combinations, Permutations and Partial Permutations*). It is often only partially observed. This is a problem in samples where factors cannot be assumed to be (1) independent or (2) in-sample. That's because (effect) observations can then be contingent on (1) the order or history of the growth process, or (2) out-of-sample factors. In practice, **two timescales determine the statistical properties of the sample growth process**: the rate ∂N_{+a} at which individual factor differences are observed (i.e., when their counterfactual effects can be observed), and the rate ∂N_{-a} at which the 'backgrounds' in which they are observed change. If the relationship between these two timescales is such that effects are observed under a large number of backgrounds, then we can be more confident about their generalisability (i.e., that effect observations will likely be reproducible in future backgrounds). On the other hand, if observations are such that different populations are observed under the same backgrounds, we can be more confident about their unconfoundedness (i.e., that effects observed reflect the same uncontrolled out-of-sample variation across all sample populations).

More precisely, the background $D_{-a} \in \mathcal{D}(X - \{a\})$ of an effect observation of factor a , is the instantaneous condition in which the effect is observed, $\Delta y[a | D_{-a}]$. We will consider definitions for $\mathcal{D}(X - \{a\})$ where it corresponds to the set of all possible values over the set of other factors, $\mathcal{D}(X - \{a\}) = X - \{a\}$, or their permutations², $\mathcal{D}(X - \{a\}) = \Pi(X - \{a\})$, and when X contains both observed and unobserved factors. For illustration, imagine all ways we can observe backgrounds across growth trajectories of the example in *Sect. Introduction*, $X = \{a, b, c, d\}$. In order to consider effect observations for a , we must 'insert' this individual factor in all its possible $3! = 6$ backgrounds observable during growth. After each insertion, we can then observe changes $\Delta y(a)$ in the outcome-of-interest, y , to understand a 's effect on y . Another way of saying this is that we need to keep a constant, while cyclically permuting all other sample factors, Fig.1(a). That is, we are called to observe the effect of a under the cyclic permutations of $d \leq m - 1$ factors present in the sample. For a factor a and sample with m factors, this defines a map¹

$$\tau(x_m) : \{b, c, \dots[m-1]\} \rightarrow \{c, \dots[m-1], b\}, \quad (2)$$

whose iteration $\tau(x_m), \tau^2(x_m), \dots, \tau^m(x_m)$ have a shifting action in the original permutation and generates the cyclic group³ of order $m-1$ [11]. Whether we require only a to be inserted in all background values with factors other than a (a single cyclic background permutation $\tau(x)$), or repeat this for all m sample factors, will change the guarantees we can make in respect to the EV and CF of effect observations. At the limit, **these two cases correspond to 1 or $m - 1$ iterations of the recursive definition of a factorial, $m! = m \times (m - 1)!$**

²where $\Pi(X)$ is the set of all permutations of the set X .

³The cyclic group generated by τ is related in an obvious way to the set of all permutations of m elements (i.e., to the S_m symmetric group), where $\tau^i(x)$, $i < m$, correspond to one partial permutation, $\tau^i(x) \in S_m$, for all⁴ $x \in \mathcal{P}(X)$ (*Sect. Combinations, Permutations and Partial Permutations*).

⁴where $\mathcal{P}(X)$ is the power-set of X and we use high caps to indicate counts over unique values.

Define, therefore, variable $N_{+x_0} = |\mathcal{P}(x_0)|$ indicating the number of unique in-sample effect observations⁴ in x_0 , and $D_{-x_0}(X) = |\mathcal{D}(X-x_0)|$ their out-of-sample backgrounds. The total number of counterfactual effect observations (i.e., all effects under different backgrounds) for the factors x_0 is then given by

$$0 < N_{+x_0} \times D_{-x_0}(X) \leq 2^{|x_0|} \times (m-|x_0|!), \quad (3)$$

(Sect. *A Representation for Sample Effect Observations*) which is large in the factorial-based definition of a 'background', $D_{-x_0}(X) \gg N_{+x_0}(X)$. A key question then becomes, as sample dimensions grow, what is the **asymptotic number of backgrounds that effects will typically be observed under?**

2.2 Unobserved-Observed Combinatorial Shuffling (Accuracy)

To illustrate how models trained in the samples with the previous characteristics can have their accuracy constrained, we can relate the previous growth process to a traditional combinatorial randomization scheme such as Gilbert–Shannon–Reeds (GSR) shuffling [10, 11]. Like before, start with a sample containing only one population $x_0 \subseteq X_m$ (set of binary attributes). Let a population with the same attributes as x_0 be represented by the string containing only 1 values (for factors a, b, c, \dots). A string containing differences can then be written as $\sigma(x_0) = (\sigma_1, \sigma_2, \sigma_3, \dots)$, where $\sigma_t \in \{0, 1\}$. An in-sample counterfactual effect observation corresponds to measuring outcomes before and after applying the operation in Eq.(2), $\tau(\sigma_t, \sigma_{t+1}, \sigma_{t+2}, \dots) = (\sigma_{t-1}, \sigma_t, \sigma_{t+1}, \dots)$. We can use the same representation for out-of-sample, or yet unobserved, factors, leading to a second sequence $\sigma_{-1}, \sigma_{-2}, \sigma_{-3}, \dots$ until we observe all factors. Sample growth can then be represented as the bi-directional string with positive values in-sample and negative out-of-sample, $\sigma = (\dots, \sigma_{-2}, \sigma_{-1}, \sigma_0, \sigma_1, \sigma_2, \dots)$. In an increasing spatial sample, the current scale corresponds to the zero-index string position. Similar to the GSR, this bi-directional string may be represented, in turn, by two real numbers $0 \leq x, z \leq 1$ as

$$x(\sigma) = \sum_{t=0}^{\infty} \sigma_{-t} 2^{-(t+1)}, \quad z(\sigma) = \sum_{t=0}^m \sigma_{t+1} 2^{-(t+1)}, \quad (4)$$

The shifting action τ for two separate strings is known as a dyadic transformation, which can be thought as a 'folding' or 'shuffling' operation between them - mapping each x to a distinct z in each iteration. In this case, the transformation is between a set of factors, x_0 , and possible background, $\partial\mathcal{D}(X-x_0)$. The right diagram in Fig.1(e) illustrates the result of a GSR shuffle for a 6-letter example with half variables observed. With each shuffle, each in-sample factor, $\{a, b, c\}$ (colored bars), effect is observed under different backgrounds (e.g., the effect of a is observed under the background of d and of b under e). We would therefore expect the generalizability (EV) of effect observations to increase with each operation. There are, however, two practical problems with this scheme. First, since each factor effect is observed in a different background, each effect observation is confounded by a different factor (e.g., the observed effect of a reflects the influence of unobserved factor d). Second, GSR shuffling only shuffles factors under a 1-step Markovian assumption, disregarding higher-order and non-additive factor interactions. Fig.1(e) illustrates an alternative 'shuffle-and-cycle' scheme, where each shuffle step is followed by m cyclic permutations $\tau, \tau^2, \dots, \tau^m$ of in-sample factors. This alternative has three advantages: (1) each factor is observed under every background, and thus effects are confounded in equal proportion across sample populations (thus allowing us, for example, to 'factor-out' these effects more easily [27]), (2) every effect is now observed for the same permutation of unobserved factors ($\{d, c, e\}$ in the figure), and thus generalizability increases at a common rate for all factors and populations, and without Markovian assumptions, (3) this is a **limiting process for every sample** (Sect. *Combinations, Permutations and Partial Permutations*).

2.3 Hyperbolic Geometry (Sample Sizes)

To better understand the relationship between sample sizes and out-of-sample performance, we need to consider what sample sizes are required to generate all ways of interleaving samples' in and out factors. Fig.1(b) outlines two equivalent sample growth patterns that achieve this ('serial', 'parallel'). In the left, each new background is interleaved with all previous effects 'serially', analogous to the previous factorial scheme. A first out-of-sample factor, $d \in X-x_0$, requires observation of $N_{+x_0}=3$ effects, $x_0=\{a, b, c\}$, a second requires $N_{+x_0}+1$ effect observations, $x_1=\{a, b, c, d\}$, etc. In the example at the right, several out-of-sample factors are interleaved with several in-sample simultaneously, analogous to a GSR shuffle. Effects for x_0 are observed, at first, under backgrounds $\{d, e, f\}$, then $\{f, d, e\}$, etc. Both strategies lead to a geometric series of distinct background for individual effects, but at different rates. Sample sizes in both cases can be described by the hyperbola,

$$\left(\frac{\omega}{\partial N_{+x_0}^2}\right) - \left(\frac{\omega}{\partial D_{-x_0}^2}\right) = 1. \quad (5)$$

The equation expresses that, for each new in-sample factor, ∂N_{+x_0} , we are able to re-measure their effects in each new unobserved background, ∂D_{-x_0} , thus increasing their generalizability (EV). The quantity ω describes, in turn, the 'speed' in which EV is expected to increase for individual in-sample populations, Fig.1(c). Sample sizes (as opposed to effect background counts) follow known exponential, $(1 - 1/m)^t \rightarrow e^{-1}$, and binomial, 2^{-1} , growth rates in these two cases, $0 < t \leq m$ ($m \gg 1$), Fig.2(b).

When all m variables are observed across the same number of backgrounds (i.e., in 'balanced' samples⁵), samples following Eq.(5) have sizes, n_{x_0} , increasing according to a Fibonacci sequence, $2 \times \partial n_{x_0} = \partial N_{+x_0} + \partial D_{-x_0}$, and thus asymptotically assuming half-golden background-to-effect rates,

$$\frac{\partial N_{+x_0}}{\partial D_{-x_0}} \rightarrow \frac{\phi}{2}. \quad (6)$$

In conclusion, these equations describe systems that permute effect observations, but whose number of effect observations are limited to under-factorial sample sizes. Samples following half-golden background-to-effect ratios, Eq.(6), observe effects across approximately the same number of backgrounds across all its populations, leading to effect estimates that have, simultaneously, increasing EV and small CF sustained throughout sample growth.

Fig.1(f) illustrates the expected relationship between maximal accuracy of samples, $ACC\left(S; X\left[x_0, d_{x_0}\right]\right)$, and sample sizes. The diagram shows sample size divided by square size vs. ACC (Sect. A Representation for Sample Effect Observations). In complete systems (top-curve), the observation of one square (full set of sample effect observations) is enough to generate accurate effect observations. In incomplete systems (bottom-curve), the number of effect observations necessary for accuracy can grow factorially with the number of unobserved factors - requiring very large samples to achieve similar levels of accuracy. This is a conservative estimate which can be abated by increases in periodicity and independence in out-of-sample factors. In real systems, where factors are typically highly correlated, it seems to be a reasonable upperbound for accuracy. We demonstrate these alternative patterns of sample growth in Sect. Results (EV vs. EV-CF) using

- Descriptive statistics (sample factor distributions, extreme-value distributions, rankings and autocorrelation),
 - Large-scale black-box prediction of economic growth (multiple state-of-the-art supervised methods),
- calculated across samples ranging from the street to national spatial level, and over 100 years of the US census.

⁵this is analogous to notions of balancedness in experimental designs [22] but require milder conditions than equal-size populations, being observable in multi-frequency and multi-scale processes, and being observed in real-world systems, as demonstrated in Sect. Results.

3 Related Work

The Shapley-value [28] has become an essential tool across disciplines to estimate the importance of variables from the output of black-box systems (i.e., whose inputs can be manipulated exhaustively at will)[21, 3, 7]. The value can be interpreted as the enumeration of all counterfactual effect observations in a **fully-observed** system. This makes the Shapley value an instance of an U-Statistic and a permutation-based statistic [20, 13]. The value $\varphi(a)$ was devised first to quantify the importance of a given player a in a m -player game, and it can be written as

$$\varphi(a) = \frac{1}{(m-1)!} \sum_{\pi \in \Pi(X-\{a\})} [y(P_a^\pi \cup \{a\}) - y(P_a^\pi)], \quad (7)$$

where $\Pi(X - \{a\})$ enumerate all permutations π of a set of size $m - 1$, y is a game utility measure, and P_a^π is a possible coalition⁶ among players (not including a), formed in the order π . Each quantity under bracket is a counterfactual observation of the effect of a (i.e., under all distinct backgrounds and their orderings). Eq.(3) counted the number of such observations for each population in a sample. The value is an ideal, as its calculation is NP-complete [5] and, when quantifying variable importance, it assumes there are no unobserved causes in the sample. Due to sample correlations this equation cannot be used, as well, with random sampling. Despite these shortcomings, the Shapley-value formulate crucial relationships among permutation of inputs when calculating sample statistics and their unbiasedness or 'fairness' [7, 3]. Calculating the **expected number of permutations that can be enumerated in samples, or locations, as proposed here, should thus sets strict bounds for their unbiasedness**, and offer a finer-grained illustration of these relationships. While the relationship between the Shapley-value and fairness of black-box predictors is known [28, 21], their relation to generalization is perhaps more surprising [27]. A quantity that becomes central to the formulation of accuracy bounds, Eq.(1), and tradeoffs between the two previous learning problems, is the growth rate, ω , in enumerable permutations across systems' spatial levels. Because the Shapley value cannot be calculated in practice, random sampling is often employed as an approximation in Shapley-based importance ranking [21]. The notion of sample squares (*Sect. A Representation for Sample Effect Observations*) was devised from the sample's set of 'observed permutations', with which effects can be calculated without assumptions of independent and identically distributed factors. Using square sampling, is advantageous not only for samples with factor correlations, but, particularly, incomplete samples, where, as formulated, assumptions of large random sampling become unrealistic due to their factorial requirements on effect observations. These gains were demonstrated in [27] and are revisited in *Sect. Results*. Furthermore, the theoretic relationships here elucidate sampling size requirements and EV-CF tradeoffs and limits for non-parametric variable importance and effect estimation.

A key element of the previous solution is that sample units in the same location share a large number of unobserved ('external') and uncontrolled factors. Studying samples with increasingly inclusive and distinct counterfactual effect observations can reveal conditions for effect generalization across samples and space. Consider the single factor case. Let $X-\{a\}$ be the set of external factors for population $\{a\}$. In a random sample with a single treatment indicator a , it follows that $\mathbf{E}\{ p[a = 1 \mid \mathcal{D}(X-\{a\})] \} = 2^{-1}$, as, at each 2 time intervals, we are expected to rebalance (random variables are bold). This is the rationale underlying, for example, Randomized Control Trials [22]. A location with this property has a single balanced population, $\{a\}$, and common external factors, $X-\{a\}$. We can alternatively say that $p[a = 1 \mid \mathcal{D}(X - \{a\})] = 0.5$, or, $a \perp (X-\{a\}) \mid \mathcal{D}(X-\{a\})$, which are typical non-confounding conditions [29, 26]. If units in the location share the same uncontrolled factors and have the same number of members with a as without a , then expected outcome differences between them correspond to a 's effect, conditional on the common variation, $\mathbf{E}[\Delta y(a) \mid x_0 = \mathcal{D}(X-\{a\})] = y(x_0+\{a\}) - y(x_0-\{a\})$. Learning systems and agents in such locations operate with fair estimates of a 's impact (albeit, with low EV). In a square, in contrast, **all m sample factors are balanced simultaneously**

⁶a player set describing a possible cooperation structure in the game with value $y(P^\pi)$ when formed.

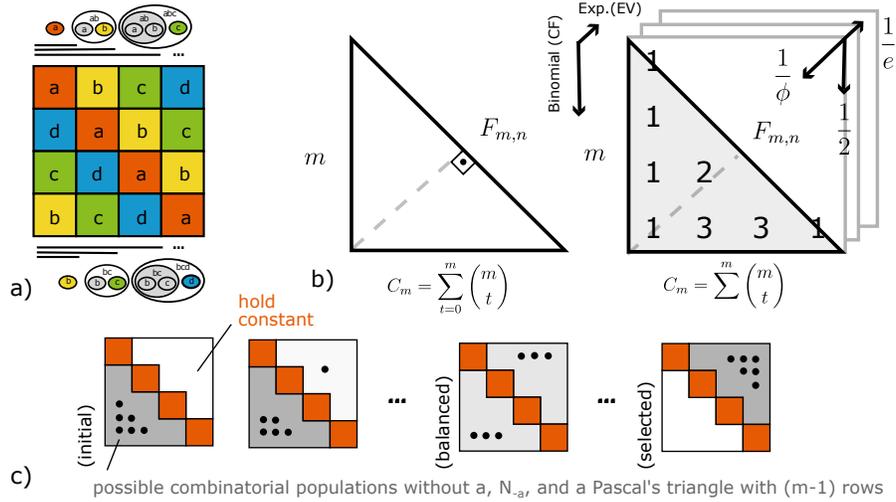


Fig. 2. (a) a $m \times m$ Latin-Square ('square') for a sample unit x_0 ($m = 4$) and the combinatorial relationships among the sample units placed across its different cells (Venn diagram, factor intersections are shown in grey and singleton differences in color), (b) Binomial ($\frac{1}{2}$), Fibonacci ($\frac{1}{\phi}$), and Exponential ($\frac{1}{e}$) rates across squares lead to hyperbolic relations among population sizes, each square's triangle altitude (dashed) is related to samples' effect-to-background rate, ω , (c) a sample population 'sweep' for factor a , where the rate of insertion of a in populations is held constant, the figure illustrate population sizes as dots and three phases of sample growth: initial (no population has a), balanced (the same number of populations have and don't have factor a), and possibly selected (where all populations have a).

($m > 1$). While single-factor balance requires a binomial series, balancing several requires Fibonacci - i.e., square 'altitude' expansion (*Sect. Sample EV-CF Growth Patterns*). Each population, in this case, follows asymptotic sample size rates $\mathbf{E}[\frac{\partial N_{+a}}{\partial N_{-a}}] = \frac{\phi}{2}$. Square accumulation thus increases the EV of all its populations simultaneously [27] - making them useful, for example, to understand sample accuracy limits across scales, Eq.(1).

4 The Combinatorial-Hyperbolic Relationship in Sample Growth

4.1 A Representation for Sample Effect Observations

The set of all counterfactuals accumulated by sample growth at one instant can be visualized with a Latin-Square ('square'), Fig.2(a). Fig.1(a) illustrates two standard ways of visualizing the complete set of $2^{|x_0|}$ in-sample effects: as the number of edges of a hypercube of dimension $|x_0|$ or sums of Pascal triangle's $|x_0|$ -th row. The square will serve, in addition, as basis for non-parametric effect estimates across sample factors. For a fixed unit or population x_0 , it represents a stratification, or 'placement', of all *other* populations, x_i , across square cells, with repetition. The completeness or incompleteness of squares, for each x_0 , will have a stipulated impact on the EV or CF of their effect observations. In particular, for m factors ($a, b, c, \dots, [m]$) the square enumerates all singleton effect observations possible from the sample's m -way effect differences. Its first column contains counterfactual effect observations for $\{a, b, c, \dots, [m]\}$ (i.e., conditioned on all other $m - 1$ factors being the same as, or 'overlapping with', x_0). The second column contains singleton effect observations possible from the previous observations (with size 1 difference and $m - 2$ intersection with x_0). These effect observations are thus conditioned on one further factor observation (i.e., on the factor in the preceding column). The third column contains singleton effect observations possible from the previous observations (size 1 difference and $m - 3$ intersection, etc.). Fig.2(a) illustrates these combinatorial patterns with Venn diagrams for each cell, where a cell's pairwise intersecting factors are grey and singleton differences are colored. This iterative procedure enumerates all possible singleton effect observations

in a sample. The square diagram shows only the singleton effects (cells), with their conditioning factors implied. Fig.1(a) illustrates that the square contains all in-sample backgrounds (each a Pascal's triangle) for a fixed factor a , and thus $2 \times 2^{m-1}$ unique effect observations of a . Each of its diagonals contains all observations for the effect of a given sample factor. Taking columns to mark time progression, the main diagonal thus marks the point of insertion of factor a (e.g., insert a in populations $\{d\}, \{d, c\}, \{d, c, b\}$). Its lower triangle records the populations without a (with size N_{-a}) and the upper triangle with a (with size N_{+a}).

The square of order $m \times m$, as a whole, contain effect observations where all factors are observed under all m -cycles of a fixed permutation (e.g., $\{a, b, c, d\}$ in Fig.2(a)). Squares of increasing orders thus captures effect observations under increasing Markovian orders (i.e., conditioned across larger times or backgrounds). The relationship of sample permutations to measurements' unbiasedness is a cornerstone of the most widely accepted Theory of Non-parametric Statistics, U-Statistics [20, 13] and of Shapley value based estimates of variable importance (*Sect. Related Work*). The relationship to generalizability has been discussed in [27], and is reviewed, and expanded, below. The full set of $\sum_{t=0}^m \binom{m}{t} = 2^m$ effect observations observable in a sample of dimension m collect 1 square for each of its sample populations, $x_i \subseteq X_m$. It suggests then a natural sample limit for the generalization of effects. Fig.1(c) illustrate the resulting phases for samples under growth with a factor a : no unit includes a (initial), as many units include a as not (balanced), and all include a (selected). Eq.(5) should hold across all such scenarios ($N_{+a}, N_{-a} > 0$).

4.2 Combinations, Permutations and Partial Permutations

The statistical concept of a 'population' is often associated with combinatorial combinations, as a set of sample units with a given attribute combination (e.g., high-income white males). There are thus $\binom{m}{t} = \frac{m!}{t!(m-t)!}$ populations of size t . A problem with this definition is that it leaves unspecified all non-population factors. To **define a population we imagine, instead, that we fix the $m - t$ population factors and vary (i.e., 'permute') all t non-population (i.e., 'external') factors**. This leads to a combinatorial structure known as a partial permutation. The number of partial permutations for a population of size t is $\binom{m}{t} \times D_{m-t}$, where D_{m-t} is the number of *derangements* (permutations without overlaps),

$$D_{m-t} = \sum_{k=0}^{m-t} \frac{(-1)^k}{k!}. \quad (8)$$

The full set of $m!$ permutations of size m , and all sample growth trajectories, can then be formulated as sets of partial permutations, using a well-known definition for factorials,

$$m! = \sum_{t=0}^m \binom{m}{t} \times D_{m-t} \quad (9)$$

$$= \left[\underbrace{\cosh(m-1) + \sinh(m-1)}_{\frac{\cosh(m-1)}{\sinh(m-1)} \xrightarrow{m} \omega} \right] \times (m-1)! + 1. \quad (10)$$

The term $C_m = \sum_{t=0}^m \binom{m}{t}$ in Eq.(9) corresponds to a single Pascal triangle and half-square (i.e., one set of all differences) for each sample population, and Eq.(9) to all squares. The number of observed permutations in a sample can thus be specified succinctly by its number of squares and their derangements. Samples with no missing variables require the observation of few derangements (no relevant exogenous variation) for accurate effect observations, while incomplete samples require the observation of many derangements [27]. The odd and even parts of Taylor's expansion of Eq.(9) leads to hyperbolic trigonometric functions, Eq.(10) (proof in

the *Supplementary Material*). They indicate the 'period' in which full sets of permutations are collected. The parametric equations for the hyperbola's right branch in (x, y) cartesian coordinates, Eq.(5), are $x = \omega \times \cosh(N_{+a})$ and $y = \omega \times \sinh(N_{-a})$, Fig.1(f). We will see that these quantities are related to in-sample effect to background sampling rates, ω . This quantity will be essential to describe statistical tradeoffs across growing samples.

4.3 Sample EV-CF Growth Patterns

According to the previous, an unbiased effect estimate for a is an average across possible effect backgrounds, and constitutes an U-Statistic [20]. There are $F_{m,n} = \sum_{t=0}^{m-1} \binom{m-t}{t}$ such sequential observations⁷. To generate all of them, we need to fix each effect observation's first, second, third, etc. factors in order. $F_{m,n}$ corresponds to the sum of the number of observations necessary to fix any first factor, $\binom{m-1}{1} = m-1$, then $\binom{m-2}{2}$ to fix a second from the remaining, etc. until all $m-1$ factors are used.

The relationship in Eq.(5) corresponds to the Cartesian equation of a rectangular hyperbola, $N_{+x_0} \times N_{-x_0} = c$, where c is constant (although well-known, this is formulated in the *Supplementary Material* for completeness). According to the previous, these two quantities have different limits, however, $N_{+x_0} \in [1, C_m]$ and $N_{-x_0} \in [1, F_{m,n}]$. The relationship can thus describe sample limits in large-populations by substituting $N_{+x_0} = C_m$ and in $N_{-x_0} = F_{m,n}$ in Eq.(5). As formulated next, the same result can be established from known rates across Pascal's triangle.

The two previous quantities, C_m and $F_{m,n}$, appear in Pascal's triangle (adjacent side and altitude), Fig.2(b). Since the main diagonal marks a 's possible 'times-of-insertion', the square's upper triangle contains the set of all counterfactuals with a , and the lower, without a , Fig.2(b). In respect to effect observations, we say that each individual effect observation is observed under $F_{m,n}$ in-sample backgrounds for each derangement D_n (or twice this value in balanced samples). The sample effect-to-background enumeration rate ω , at time t , is thus defined as $\omega = \frac{F_{m,n}}{D_n}(t)$, or, the number of in-sample background observations, $F_{m,n}(t)$, per derangement, $D_n(t)$, across all populations in the sample.

The growth of C_m , D_n and $F_{m,n}$ for N_{+a} or N_{-a} assume Pythagorean relations⁸, Fig.2(b),

$$\left(\frac{\partial C_m}{\partial D_n}\right)^{-2} + \left(\frac{1}{\partial D_n}\right)^{-2} = \left(\frac{\partial F_{m,n}}{\partial D_n}\right)^{-2}, \quad (11)$$

Eq.(11) suggests the visualization of sample growth as hyperbolae⁹ with increasing radius D_n , Fig.1(f). In this limiting expression of Eq.(5), C_m corresponds to all possible individual sample populations, N_{+x_0} and N_{-x_0} , and F_m to in-sample effect observation backgrounds. Fig.1(f) shows the hyperbolic asymptotes $C_m = F_{m,n}$ and $C_m = -F_{m,n}$ (dashed). They represent growth with constant EV, $\partial D_n = 0$. The figure also shows the asymptotic sample (vertical black line) where exactly all observations have factor a , $F_{m,n} = 0$. **Under this condition, no estimator, algorithm, or agent is able to estimate a 's effect non-parametrically.** It represents the sample with minimum EV, while outward hyperbolae, samples with increasing EV. Unique background count growth in this direction follows a Fibonacci series, whose rate is the Golden number. It is well known that the rows, columns and diagonal of Pascal's triangle are associated with binomial, exponential and fibonacian rates. Notice then that $\frac{\partial D_n}{\partial C_m} \in [1/e, 1/2]$, as growth can range between $\frac{\partial C_m}{\partial m} = 2$, and $\frac{\partial D_n}{\partial n} = 1/e$, Fig.2(b). The first is due to $C_m = 2^m$, and the second was famously established by Euler [30]. In the previous nomenclature, the first is associated with balanced or Unconfounded (CF) growth, and the second with EV sample growth. The golden ratio is associated, in contrast, with high-dimensional balanced growth of samples and populations, EV-CF growth, and with squares, Eq.(11).

⁷ $\binom{m-t}{t} = 0$, when $t > m$.

⁸the equation uses the Pythagorean theorem in its reciprocal form, as it includes the triangle's altitude.

⁹the equation for a hyperbole is $(\frac{x}{a})^2 - (\frac{y}{b})^2 = r$, with a and b its vertices and r radius.

More specifically, squares are associated with the assumption that $\frac{\partial C_m}{\partial F_{m,n}}$ is constant across factors (i.e., hyperbolae with constant radius), Eq.(11). It indicates that factors' diagonals are the same size, and the population structure is, overall, a 'square'. Finally, the known hyperbolic relationships

$$\tanh(n) = \frac{\sinh(n)}{\cosh(n)} = \sqrt{1 + \left(\frac{\partial D_n}{\partial C_m}\right)^2} = \omega^{-1}, \quad (12)$$

suggest expressing sample effect-to-background enumeration rates ω in terms of $\tanh(n)$ ¹⁰. Also note that this definition for ω coincides with that of Lorentz factor γ [8, 14], best known as a time correction between frames-of-reference in the physical sciences. Here, it preserves frequency relations among factors, N_{+a}/N_{-a} , under changes of basis of the type $x=x_0+\{a\}$ and $x=x_0-\{a\}$. As suggested by Borel [4], it is natural to think of the transformation as a hyperbolic rotation (analogously to the typical trigonometric). We will illustrate many of these mathematical abstractions using real-world spatial data in *Sect. Results*.

5 Results

We will now illustrate the formulated combinatorial and statistical generalizability limits in an important real-world problem: out-of-sample economic growth prediction across increasing spatial extensions (i.e., samples with increasing census individuals). Data used encompasses microdata of American decennial censuses from 1840 to 1940, and approximately 65 billion individual-level records. This time range corresponds to the decades of American urbanization. We consider the economic and demographic changes as we go, spatially, from the household spatial-level, d_0 in lat-lon distances, all the way to the national level, for each studied year. We thus create samples with units at arithmetically increasing levels, $d_{t+1}=d_t+\Delta d$ (starting from d_0). We repeat this for approximately 60K American locations, x_0 . Each full spatial analysis is then reproduced independently across years (thus avoiding issues related to extended longitudinal data). Fig.3(b, top-right) shows two locations in New York City, which share a large amount of external variation (i.e., economic and demographic variations across the rest of the country). The resulting nation-wide transversal captures combinatorial patterns of populations' differences and overlaps in samples, for all x_0 , as we increase scale. Our main goal is to illustrate how, consequently, **generalizability change across spatial-scales**, according to the stipulated model and limits. We first study sample correlations and sizes, demonstrating they follow the previous hyperbolic relationships. We then repeat previous out-of-sample prediction tasks with this new census data and increasing spatial levels - thus adding to previous evidence presented for a combinatorial counterfactual model for sample generalizability [27].

5.1 Descriptive Statistics (Sample Sizes and Correlations)

We illustrate the consequences of Eq.(12) to sample properties using Autocorrelation functions (correlations) and hyperbolic co-tangent (coth) regressions (sample sizes) in large-scale census data. These considerations will be key to solving our main problem, Eq.(1), as the accuracy of agents and algorithms operating in samples are directly determined by sample sizes and their combinatorial patterns [27]. Economic distribution across space can be described by the primary occupation and industry of all census individuals [1, 16] (e.g., 'carpenter' or 'executive assistant'). We start with this set of variables, and discuss the full set of variables, including non-economic, in the next section. Fig.3(c) illustrates empirical frequencies for all occupations (each a curve) at 4 different spatial-levels in Massachusetts (MA) and New York (NY), 1880. They were the country's economic centers until the 19th century. The distribution has the familiar shape of a wave that moves to the left. New York reaches a stationary shape at a lower level d_{sq} . We demonstrate these correspond to levels where squares are completed across factors.

¹⁰with $n = \operatorname{arctanh}(\omega^{-1}) = \operatorname{arctanh}\left(\frac{D_n}{F_{m,n}}\right)$, which, lets n be the number of accumulated derangements per fixed $F_{m,n}$ (i.e., per square), as expected.

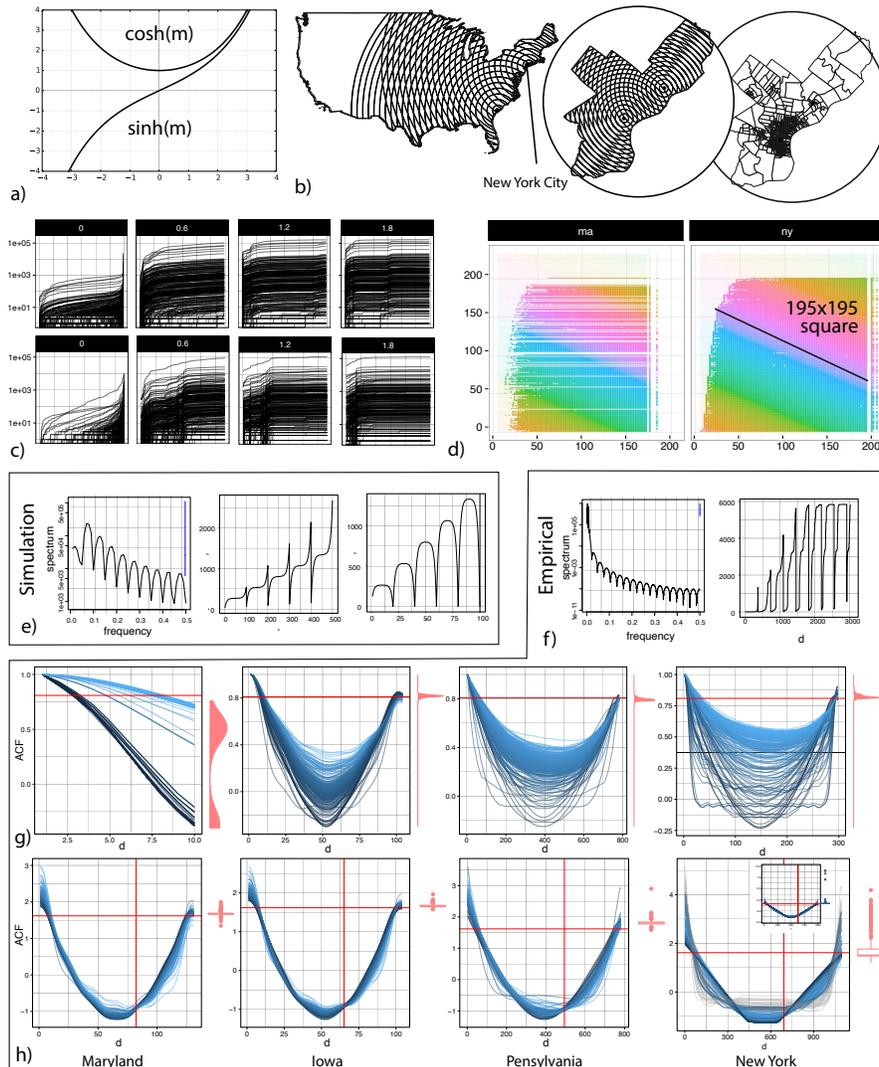


Figure 3. (a) sinh and cosh functions, (b) increasing spatial-levels at two example locations (national and city-levels), (rightmost panel) finest spatial-level for New-York City, (c) occupation frequency ranks vs. location across 4 example scales, each curve is an occupation, (d) enumerated Latin-Squares histograms for Massachusetts and New York, the latter has a square with almost all occupations, (e) periodogram of $cosh(100)+sinh(100)$, $sinh(100)$ and $cosh(100)$, (f) per-occupation periodogram and series example from (c), (g) auto-correlation vs. spatial-level trace catenaries (free-hanging ropes) for each occupation (1880), probability distribution of their slack (red, sidepanel) indicate a fixed ω per factor at 0.81 (red horizontal line), Eq.(13), (h) standardized catenaries across *all* years, boxplots (red, sidepanel) show slack invariance and constant ratio between sinh and cosh growth for all locations, years and occupations, $m \times (1 - \frac{1}{e})$ (red vertical line) is a fixed point in binomial-exponential (EV-CF) to exponential (EV) rate transitions.

All squares in a location can be enumerated through an expensive computational procedure [27]. Fig.3(d) shows histograms, where each color corresponds to one of 220 occupations. NY has a spatial square that extends to almost all occupations, while MA has missing occupations (horizontal gaps) in comparison.

5.1.1 A 'Hanging-Rope' Model for Unbiased Sample Growth. The Catenary is a hyperbolic curve with long scientific history and describes a free-hanging rope [9]. Unlike circles and geodesics, they are sums of exponentials. Their equation in (x, y) Cartesian coordinates is $y = \cosh(x)$, and their length is $l = \sinh(x)$, making them useful to demonstrate the previous model, Eq.(12), and increases in enumerable permutations across spatial levels. We demonstrate that both the observed shape, Fig.3(h), and parameters, Fig.3(h, boxplots), of spatial correlations follow predictions from the previous model. Before considering catenaries, however, Fig.3(a,e) illustrate the overall shape of the previous hyperbolic functions, and their frequency-based representations. Fig.3(a) depicts $\cosh(n)$ and $\sinh(n)$, and, Fig.3(e) the periodogram of $\cosh(n)+\sinh(n)$, $\sinh(n)$ and $\cosh(n)$ where $n = 100$. Fig.3(f) illustrates the *empirical* periodogram of curves in Fig.3(c), which resemble the simulated.

Fig.3(g) shows auto-correlation function (ACF) for all locations across 5K spatial-levels (as those illustrated in Fig.3(c)), until the state level (*Sect. Methods, Supplementary Material*). They trace catenaries. The horizontal line $y = 1.0$, of unitary correlation, is associated with the limit $F_{m,n} = 0$ where, despite the increasing scale, no population differences are added to the increasing samples. Each single catenary is a set of samples with constant $C_m/F_{m,n}$, which is a **defining property of squares**, Eq.(11). Fig.3(g) illustrates 4 typical cases among states. Plots for all states are available in the Supplementary Material. Maryland has linear decreases in auto-correlation. From 1840, the USA economy and cities become increasingly interdependent. After 1900, no longer any state had such linear correlation signatures. Periodic and linear (zig-zag) auto-correlations, with period $m/2$, are related to non-increasing EV, Fig.1(f, black vertical line). Periodic and exponential correlations, without growth, correspond to catenaries with $h = 0$ (where a system returns to its original state after a lag). The defining characteristic of the catenary is $\partial y/\partial x = l/h$, where l is its length and h its 'slack', or, difference in height, y , between its two hanging points. Standardizing catenaries [9] (i.e., making l unitary and h constant)¹¹ thus makes its slack h indicate ω during sample growth,

$$h = \omega \approx \frac{\phi}{2}, \quad (13)$$

which, according to Eq.(6), should assume half-golden values for small D_{-x_0} . Fig.3(h) shows standardized catenaries for all years and locations. It indicates that \tanh per factor remains constant across a range of levels, up to d_{sq} , starting at the local. This was anticipated by Eq.(12). The rate, up to d_{sq} , is 81% of correlation. Fig.3(h, sidepanel, red) shows **box-plots** for h , across all levels, years, occupations, and locations. For all spatial-levels below d_{sq} , factors remain balanced, with binary-exponential rates (i.e., hyperbolic functions with period $m/2$). Levels above d_{sq} reverse to exponential growth. We called this a transition between EV-CF and EV growth. This is indicated in plots by the dislocation of the catenary center from $m/2$ to $m(1-1/e)$ (red vertical lines). We reproduce the same results, Fig.3(g,h), with standard **Pareto regressions** in *Sect. Methods* of the *Supplementary Material* - as alternative to these graphical depictions. Fig.4(d) shows estimated levels d_{sq} for all states, across years. Levels d_{sq} converge across years for all states. New York has 2-level squares. Fig.3(h, upright-panel) shows catenaries for its lower-level square, and Fig.4(b) illustrate levels cartographically. The two squares' factors are disjoint (gray, lower factors taking exponential rates in higher). American states have had through their histories very different work forces and regional distributions. While catenary lengths are different across occupations, Fig.3(g), their slack (and \cosh - \sinh growth ratio) **remains invariant across all locations, years and occupations**, Fig.3(h). These plots thus graphically illustrate the previous combinatorial constraints for the unbiasedness and predictiveness of learning systems across space. We return to this discussion in *Sect. Predictive Statistics*.

5.1.2 Permutations in Heterogeneous Samples. Zipf's law and distribution are central to the study of city size distributions [24, 12, 2]. The law is based on a frequency ranking of studied factors, and thus, on one of their

¹¹ $\partial y/\partial x = \partial \cosh(x)/\partial \sinh(x) = \sinh(x)/\cosh(x) = \tanh(x)$.

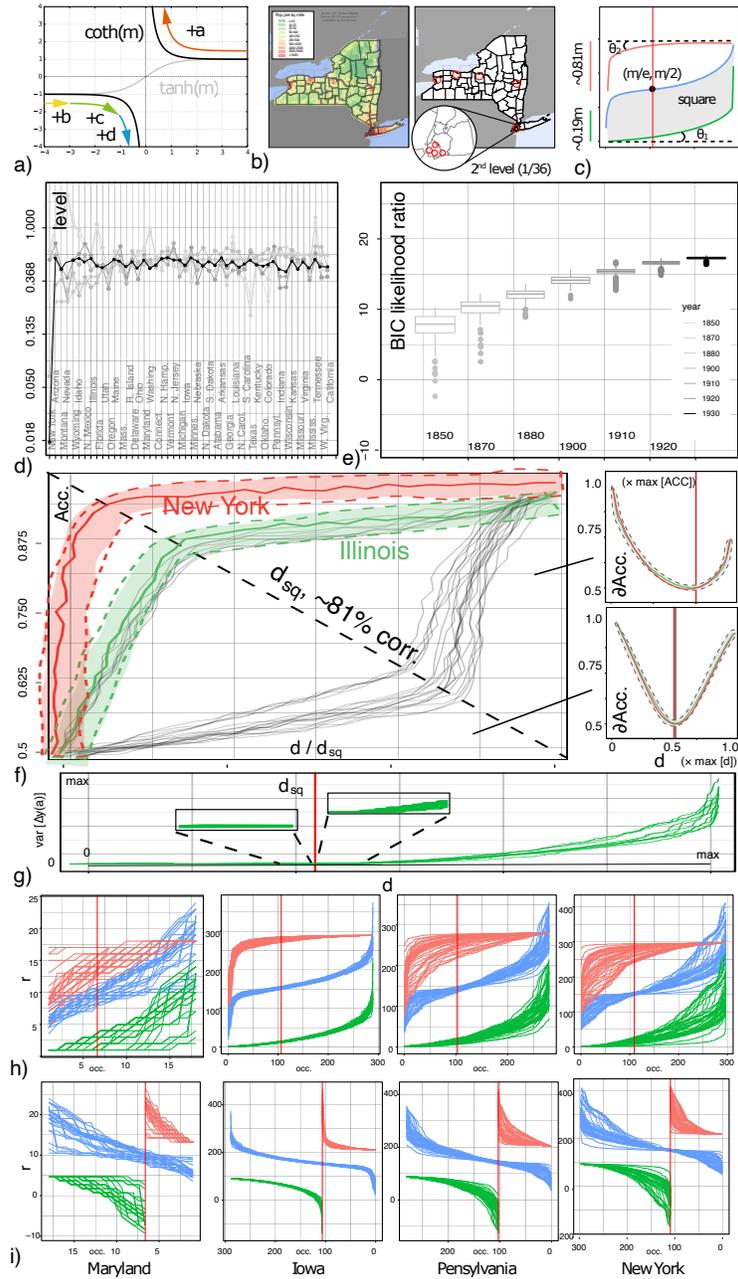


Figure 4. (a) coth and tanh functions, colored arrows illustrate square factor frequency increases with scale, (b) New York (NY) state population density (left), NY has squares at two levels d_{sq} , at 0.018 and 0.65 lat-lon distances (red circles), (c) schematic depiction of frequency rank vs. spatial-scale plots in (h), (d) d_{sq} (spatial-level of first square) for states and years, (e) BIC likelihood of coth over a Zipf model, (f) state-of-the-art growth prediction's accuracy with increasing spatial-level d , d_{sq} is diagonal (dashed), (g) min. (green) and max. (red) frequency ranks across locations, each curve is a scale, blue curves indicate square size, which follows a tanh function, (h) coth model, as illustrated in (a), with empirical data.

permutations. It is, here, associated with homogeneous samples (i.e., samples with little across-factor variation). Fig.4(a) depicts the overall shapes of $\tanh(n)$ and $\coth(n)$ functions. Fig.4(h) shows occupations' minimum frequency rank, r_0 (green), across all locations in increasing spatial levels, as well as their maximum rank, r_ω (red). The former is the minimum frequency ranking order of one given occupation across all the level's locations. The latter is the maximum (these are formulated explicitly in *Sect. Methods*). The latter is related to Zipf's frequency rankings and the Pareto distribution (*Sect. Methods*), as the three are Power-laws. Each curve in the figure corresponds to one spatial-level and occupation. With a homogeneous sample, we expect one highest-rank industry across all locations, and thus $r_\omega - r_0 = 1$. What we observe, however, is that factors are ranked in constant-sized ranges, as visualized in squares. Each factor is the highest ranked in *some* location, the second in other, etc. These rankings define an arithmetic series - $r_0, r_0 + 1, r_0 + 2, \dots, r_\omega$ - for each factor. The series has mean $\bar{r} = \frac{r_0+r_\omega}{2}$, which is also shown (blue). The previous model predicts both that $r_\omega - r_0$ is constant, and that it reflects the enumeration rate ω . Fig.4(h) shows that empirical rankings have constant $r_\omega - r_0$, with increasing r_0 . A closer examination of both branches (red and green) reveals they correspond to the positive and negative sections of the $\coth(n) = 1/\tanh(n)$ function, Fig.4(i,a).

Imagine the following process: pick a location x_0 , and its most and least-frequent factors (i.e., with rank 1 and m). Label them, respectively, a and z . Balance z to match a 's frequency. Move one spatial-level up, pick another z , balance, and repeat. This is the process described by Eq.(5). Each square row corresponds to a single derangement and background, where $\partial n_{+a} = \omega \times \partial D_a$ is the number of units in cell a . The cost to balance each z is thus n_{+a}/ω per row. For all locations x_0 , and levels $d_0 \leq d \leq d_{sq}$,

$$\begin{aligned} n_{+a} \times \frac{1}{\omega} - n_{+z \in X - \{a\}} &= 0, \\ n_{+a} - n_{+z \in X - \{a\}} \times \coth(n) &= 0. \end{aligned} \tag{14}$$

The \coth function has the interesting property of **separating, by sign, each location's 'background' and 'effect' phases**, and describe more directly how squares are completed. This is illustrated in Fig.4(a) as one hyperbolic rotation, with subsequent square derangements leading to others.

Methodologically, this suggests fitting a \coth function to observed frequency ranks. A Zipf-distribution can be fit by Power-law or Pareto distribution regressions (*Sect. Methods*). Enumeration rate increases imply increasing permutations - and thus differences between min. and max-frequency ranks. This predicts that Zipf-Pareto regressions will become increasingly inaccurate (compared to \coth), as cities become more heterogeneous. Fig.4(e) shows increase of up to 18 times fit likelihood favoring the \coth model throughout the studied period, according to a Bayesian Information Criterion.

5.2 Predictive Statistics

What impact does the presence of squares in samples have statistically (in respect to bounds to their predictiveness and biasedness)? Fig.4(f) demonstrates a result, using census microdata, with an Accuracy vs. Spatial-level plot (see [27] for others). Samples in the previous section contained sample units' primary occupations [6, 25]. This led to binary samples of dimension $m = 543$ (and each unit seen as a 543 length binary vector). For Fig.4(f), all variables in the American micro census were, instead, used [17]. Each census binary variable lead to one field, each categorical variable to as many binary variables as the size of their domain (as defined by the census) and continuous variables to 8-bit vectors (corresponding to their 8 quantiles). The final sample had 10.055 variables, including information on a broad range of population characteristics, including fertility, nuptiality, life-course transitions, immigration, internal migration, labor-force participation, occupational structure, education, ethnicity, and household composition. These variables can be correlated, colinear and spurious. Each of the

multiple state-of-the-art classifiers employed next will deal with these statistical pitfalls in their own proposed ways.

The classification task in Fig.4(f) is to predict whether a given occupation will grow (enlist further members) in the next time interval (10 years ahead), for each location x_0 . Detailed description of algorithms used, and their hyperparameter optimization, can be found on [27]. They include Neural Network Models, Generalized Linear Models, Boosting Models, Generalized Additive Models, Random Trees, LASSO and Ridge regressions, ANOVA, Support Vector Machine, and stacked meta-learners for all previous algorithms. Spatial levels (and aggregated data) ranged from the local to the national. One million location and year were chosen randomly, each leading to a full set of spatially growing samples. The figure thus shows the maximum accuracy of 24 state-of-the-art supervised algorithms, to predict whether a given occupation will grow, or not, in a location, as we use data from increasing spatial-levels (starting with the local and reaching all national). Accuracy is defined as the number of accurately classified observations in the held-out sample. Spatial levels d_{sq} for each state are mapped to the diagonal (dashed) in the figure, and each state is a curve. The way accuracy changes across locations largely follow the shape expected by Fig.1(f). Accuracy was averaged across same-state locations to generate these curves. Bootstrap accuracy variation bands (across states' locations) are shown for the two most accurate states, New York and Illinois. We observe that New York gains little from external data, above d_{sq} , as it already contains, within its boundaries, high levels of variation. This also implies that, without unobservables, $\sim 81\%$ of the sample is sufficient for prediction. Homogeneous locations, in contrast, have incomplete squares, and observed predictions are susceptible to external and unobserved variation.

The right panels in Fig.4(f) shows the *increase* in accuracy, $\partial ACC/\partial d$, of samples encompassing increasing distances (at 0.05 lat-lon intervals, normalized over their spatial and accuracy ranges) for New York and Illinois. It is estimated by the differences in accuracy of models trained at a level d and its predecessor. The top panel shows accuracy of algorithms with samples with observations $d > d_{sq}$, and bottom panel $d \leq d_{sq}$, across all locations in those states (gray ribbons show their standard deviation). These patterned differences in accuracy change in the output of supervised black-box algorithms mirror the shape of functions in Fig.3(g,h). This is expected as the accuracy of systems with characteristics described above increase with pairwise correlations, Fig.2(a). The functional form for accuracy $\mathcal{F}(d, \mu) = \mu \times \sinh(d)$ take, therefore, hyperbolic forms with distinct parameters¹², $\mu = 0.5 \times m$ and $\mu = (1 - 1/e) \times m$, and constant \tanh , $0 \leq \mathcal{F}(d, \mu) \leq 1$. The 80-20 ratios, observed for correlations in Sect. A 'Hanging-Rope' Model for Unbiased Sample Growth, are thus also observed in the outcome of black-box predictors, Fig.4(f). Like before, the functional $\mathcal{F}(d, \mu)$ is an apt description in these systems only because \tanh remain constant throughout them (i.e., in squares and balanced samples), Eq.(11).

Given the balance of samples with $d < d_{sq}$, it is expected for effect estimation to be accurate in these samples, in contrast to samples with $d > d_{sq}$. In [27], we use multiple simulated scenarios to show that samples having combinatorial conditions like those with $d < d_{sq}$ facilitate causal effect estimation. This is possible because, there, we have ground truth information for the effect of variables. We do not have ground truth in this large real-world system, but the central claim here is that in samples with $d > d_{sq}$ the problem becomes precipitously harder. Fig.4(g) shows variance in effect estimation for the popular Shapley-based effect estimation [21] across all locations x_0 and distances d , in the previous samples. There is a discontinuity, and significant increase in uncertainty over effect estimates above d_{sq} (and very little below), across all locations. Together, the previous considerations suggest constraints, described by $\mathcal{F}(d, \mu)$, for supervised prediction and effect estimation in spatial systems, Eq.(1). Functions $\mathcal{F}(d, (1-1/e)m)$ and $\mathcal{F}(d, 0.5m)$ describe, respectively, spatial growth patterns during externally-valid (EV) and unconfounded (EV-CF) samples.

¹²the same is expected from $\frac{d}{dx} \sinh x = \cosh x$ and results in Sect. A 'Hanging-Rope' Model for Unbiased Sample Growth.

6 Conclusion

We studied applications of concepts from non-parametric and counterfactual statistics to sample growth processes; common, for example, in the study of spatial systems. We highlighted sample conditions where m populations can have their effect observations remain unbiased while, at the same time, increasing in generalizability. Hyperbolic functions offered a natural description and visualization for these alternative growth patterns. We used them to formulate, in particular, the asymptotic number of backgrounds that effects are observed under in large-scale data, and thus limits to effect generalizability in these data. Increase in generalizability requires exponential sample size growth. Increase with unbiasedness requires Fibonacci, with a half-golden growth ratio. Future work will describe whether the in and out-of-sample variable shuffling models described here can also lead to finer-grain and non-asymptotic limits to effect generalization based on background randomization. We demonstrated the model empirically (functional-form $\mathcal{F}(d_{x_0})$, enumerative and combinatorial properties, 3 predicted rates), and connected sample growth to the statistical environment (biases and predictability) it creates for its populations.

Acknowledgments

Funding provided by the Sao Paulo Research Foundation (FAPESP). The datasets analyzed are available in the IPUMS repository[17].

References

- [1] Pierre Alexane Balland, Cristian Jara-Figueroa, Sergio G Petralia, Mathieu P. A Steijn, David L Rigby, and César A Hidalgo. “Complex economic activities concentrate in large cities”. In: *Nature human behaviour* 4.3 (2020), pp. 248–254. DOI: [10.1038/s41562-019-0803-3](https://doi.org/10.1038/s41562-019-0803-3).
- [2] Marcus Berliant and Axel H. Watanabe. “A scale-free transportation network explains the city-size distribution”. In: *Quantitative Economics* 9.3 (2018), pp. 1419–1451. DOI: <https://doi.org/10.3982/QE619>. URL: <https://doi.org/10.3982/QE619>.
- [3] Paul de Boer and João F. D Rodrigues. “Decomposition analysis: when to use which method?” In: *Economic systems research* 32.1 (2020), pp. 1–28. DOI: [10.1080/09535314.2019.1652571](https://doi.org/10.1080/09535314.2019.1652571).
- [4] Emile Borel. *Introduction geometrique a quelques theories physiques*. Paris, 1914.
- [5] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suci. “On the Tractability of SHAP Explanations”. In: *Journal of Artificial Intelligence Research (JAIR)* (2020). DOI: <https://doi.org/10.1613/jair.1.13283>. URL: <http://starai.cs.ucla.edu/papers/VdBJAIR21.pdf>.
- [6] Census Bureau. *Alphabetical index of industries and occupations. 1950 census of population*. Washington, 1951.
- [7] Nadia Burkart and Marco F. Huber. “A Survey on the Explainability of Supervised Machine Learning”. In: *J. Artif. Int. Res.* 70 (2021), pp. 245–317. ISSN: 1076-9757. DOI: [10.1613/jair.1.12228](https://doi.org/10.1613/jair.1.12228). URL: <https://doi.org/10.1613/jair.1.12228>.
- [8] Sean Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Benjamin Cummings, 2003. ISBN: 0805387323.
- [9] Paul Cella. “Reexamining the Catenary”. In: *The College mathematics journal* 30.5 (1999), pp. 391–393. DOI: [10.1080/07468342.1999.11974093](https://doi.org/10.1080/07468342.1999.11974093).
- [10] Persi Diaconis. *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11. Hayward, CA: Institute of Mathematical Statistics, 1988, pp. vi+198. ISBN: 0-940600-14-5. URL: <http://projecteuclid.org/euclid.lnms/1215467407>.
- [11] Persi Diaconis, R. L Graham, and William M Kantor. “The mathematics of perfect shuffles”. In: *Advances in Applied Mathematics* 4.2 (1983), pp. 175–196. DOI: [https://doi.org/10.1016/0196-8858\(83\)90009-X](https://doi.org/10.1016/0196-8858(83)90009-X). URL: <https://www.sciencedirect.com/science/article/pii/019688588390009X>.

- [12] X Gabaix. “Zipf’s Law for Cities: An Explanation”. In: *The Quarterly journal of economics* 114.3 (1999), pp. 739–767. DOI: [10.1162/003355399556133](https://doi.org/10.1162/003355399556133).
- [13] Wassily Hoeffding. “A Class of Statistics with Asymptotically Normal Distribution”. In: *The Annals of mathematical statistics* 19.3 (1948), pp. 293–325. DOI: [10.1214/aoms/1177730196](https://doi.org/10.1214/aoms/1177730196).
- [14] Joseph Hucks. “Hyperbolic complex structures in physics”. In: *Journal of mathematical physics* 34.12 (1993), pp. 5986–6008. DOI: [10.1063/1.530244](https://doi.org/10.1063/1.530244).
- [15] Daniela Incezan and Luis I. Prádanos. “Viewpoint: A Critical View on Smart Cities and AI”. In: *J. Artif. Int. Res.* 60.1 (2017), pp. 681–686. ISSN: 1076-9757.
- [16] Hong Inho, Frank Morgan R., Rahwan Iyad, Jung Woo-Sung, and Youn Hyejin. “The universal pathway to innovative urban economies”. In: *Science Advances* 6.34 (), eaba4934. DOI: [10.1126/sciadv.aba4934](https://doi.org/10.1126/sciadv.aba4934). URL: <https://doi.org/10.1126/sciadv.aba4934>.
- [17] IPUMS. *U.S. Individual-level Census (United States Bureau of the Census)*. 2022. URL: https://usa.ipums.org/usa/complete_count.shtml.
- [18] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. “Prediction Policy Problems”. In: *The American economic review* 105.5 (2015), pp. 491–495. DOI: [10.1257/aer.p20151023](https://doi.org/10.1257/aer.p20151023).
- [19] Christopher Krapu, Robert Stewart, and Amy Rose. “A Review of Bayesian Networks for Spatial Data”. In: *ACM Trans. Spatial Algorithms Syst.* (2022). Just Accepted. ISSN: 2374-0353. DOI: [10.1145/3516523](https://doi.org/10.1145/3516523). URL: <https://doi.org/10.1145/3516523>.
- [20] A. J. Lee. *U-statistics : theory and practice*. New York: M. Dekker, 1990. ISBN: 0824782534.
- [21] 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.
- [22] Douglas C Montgomery. *Design and analysis of experiments*. New York: John Wiley, 2001. ISBN: 0471316490; 9780471316497.
- [23] Stephen L Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press, 2007. ISBN: 0521671930; 9780521856157; 9780521671934; 0521856159. DOI: [10.1017/CBO9780511804564](https://doi.org/10.1017/CBO9780511804564).
- [24] MEJ Newman. “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary Physics* 46.5 (Sept. 2005), pp. 323–351. DOI: [10.1080/00107510500052444](https://doi.org/10.1080/00107510500052444).
- [25] Anastasiya M Osborne, United States. Bureau of Labor Statistics. Office of Productivity, Technology, and Peter B Meyer. *Proposed category system for 1960-2000 census occupations*. Washington, D.C.]: U.S. Dept. of Labor, Bureau of Labor Statistics, Office of Productivity and Technology, 2005.
- [26] Hans Reichenbach. *The direction of time*, Berkeley: University of California Press, 1956.
- [27] 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](https://doi.org/10.1038/s41598-024-80839-8). URL: <https://doi.org/10.1038/s41598-024-80839-8>.
- [28] Alvin E. Roth. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge: Cambridge University Press, 1988. ISBN: 9780521361774. DOI: [DOI:10.1017/CBO9780511528446](https://doi.org/10.1017/CBO9780511528446). URL: <https://www.cambridge.org/core/books/shapley-value/D3829B63B5C3108EFB62C4009E2B966E>.
- [29] Donald B Rubin. “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions”. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331. DOI: [10.1198/016214504000001880](https://doi.org/10.1198/016214504000001880).
- [30] Charles Edward Sandifer. *How Euler did it (Chapter 17, pg.103)*. eng. The MAA tercentenary Euler celebration ; v.3. Washington, DC]: Mathematical Association of America, 2007. ISBN: 9780883855638.
- [31] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. “Toward Causal Representation Learning”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634. DOI: [10.1109/JPROC.2021.3058954](https://doi.org/10.1109/JPROC.2021.3058954).

- [32] Shashi Shekhar. “What is Special about Spatial Data Science and Geo-AI?” In: *33rd International Conference on Scientific and Statistical Database Management*. SSDBM 2021. Tampa, FL, USA: Association for Computing Machinery, 2021, p. 271. ISBN: 9781450384131. DOI: [10.1145/3468791.3472263](https://doi.org/10.1145/3468791.3472263). URL: <https://doi.org/10.1145/3468791.3472263>.
- [33] Shashi Shekhar and Pamela Vold. *MIT Press Essential Knowledge series*. Cambridge: The MIT Press, 2020. ISBN: 0-262-35681-3.
- [34] Ilaria Tiddi and Stefan Schlobach. “Knowledge graphs as tools for explainable machine learning: A survey”. In: *Artificial Intelligence* 302 (2022), p. 103627. DOI: <https://doi.org/10.1016/j.artint.2021.103627>. URL: <https://www.sciencedirect.com/science/article/pii/S0004370221001788>.

Received 14 November 2023; accepted 26 July 2024