

Usage of Distribution Extents in Predictive Modeling

Alex Zolotovitski

Microsoft Corporation, Redmond, WA, USA, alexzol@microsoft.com

Abstract

Distribution Extents (DE) of order k for a sample $\{x_1, x_2, \dots, x_n\}$ of a non-negative stochastic variable X can be defined as

$$E_k = \begin{cases} (\sum_{i=1}^n p_i^k)^{\frac{1}{1-k}} & \text{if } k \neq 1, \\ \exp(-\sum_{i=1}^n p_i \ln p_i) & \text{if } k = 1 \end{cases},$$

where $p_i = x_i / \sum_{i=1}^n x_i$,

and are useful measures for a number of “large values” in the sample. They were introduced by Timo Koski and Lars-Erik Persson in 2003 [1], who generalized results of L.L. Campbell [2], and are generalization of inverse Herfindahl-Hirschman Index (HHI), a commonly accepted measure of market concentration in economics, Simpson’s diversity index used in ecology and are closely related to Shannon-Wiener Index and the Rényi entropy and divergence.

In this work we describe general properties of E_k and use it in analysis of a web advertisement network, where actors are advertisers, publishers, and users, for three purposes: 1) as cut off parameters to present the network as a graph to visualize the network and to use graph theory methods 2) as independent variable in predictive modeling, and 3) as a criterion for optimization of some parameters of models.

Keywords: distribution; entropy; web advertisement network; predictive modeling.

1. Introduction

A typical question in many data mining tasks is:

We have n objects, characterized by a variable $x_i, i = 1, 2, \dots, n$ and number n is too large to consider all objects.

For simplicity we suppose that $x_1 \geq x_2 \geq \dots, x_n$. We need truncate data dropping small objects and keeping in analysis only large object. What could be criteria to choose a cut-off parameter x_0 or threshold for x_i , if we are going to keep only $x_i > x_0$?

How many large objects do we have?

Examples:

Countries: How many countries with large population? With large GDP?

Enterprises: In wide use is inverse Herfindahl-Hirschman Index (HHI), a commonly accepted measure of market concentration in economics[3]. In this case x_i - revenue,

$p_i = \frac{x_i}{\sum x_i}$ - market shares:

$$HHI = \sum(p_i^2);$$

$$InverseHHI = 1/\sum(p_i^2) \quad (1)$$

2. Properties of Inverse Herfindahl-Hirschman Index (HHI)

The inverse HHI has following practically important properties:

$$1 \leq InverseHHI \leq n;$$

$$InverseHHI = 1, \text{ when } x_1 = A, x_i = 0, \quad (1a)$$

$$i = 2, \dots, n, A > 0$$

$$InverseHHI = n, \text{ when } x_1 = x_2 = \dots = x_n \quad (1b)$$

Let $x_i = A, i = 1, 2, \dots, n_{Large};$

$x_i = a, i = n_{Large} + 1, \dots, n; \text{ where } 0 < a \ll A.$

$$\text{Then } InverseHHI \approx n_{Large} + O\left(\frac{n_{Small} a}{A}\right) \quad (2)$$

The last property (2) is the most important. It shows that if the set $\{x_i\}$ has n_{Large} large values and n_{Small} small values, then $InverseHHI$ is a little bit larger than n_{Large} . Similar statistics that is used in ecology is a Simpson's diversity index[4].

3. Generalization of Inverse Herfindahl-Hirschman Index

In (1) instead of power 2 we can use arbitrary $k > 0$:

$$E_k = \begin{cases} (\sum_{i=1}^n p_i^k)^{\frac{1}{1-k}} & \text{if } k \neq 1, \\ \exp(-\sum_{i=1}^n p_i \ln p_i) & \text{if } k = 1 \end{cases}, \quad (3)$$

where $p_i = x_i / \sum_{i=1}^n x_i$.

The second equation in (3) for case $k = 1$ is chosen to have E_k continuous at $k = 1$. Equations (3) define generalized exponential entropy(DE) of order k for a sample $\{x_1, x_2, \dots, x_n\}$ [1], that in special case $k = 1$ is ordinary exponential entropy (Shannon index) [5], and in special case $k = 2$ we get $E_2 = InverseHHI$.

We can interpret p_i as probabilities, e.g. in case of HHI, where p_i are market shares of enterprises revenues, the p_i could be considered as probability that randomly chosen dollar of revenue belongs to enterprise i .

In frame of the probability interpretation, equations (3) define generalized exponential entropy, or distribution extents (DE) of order k for a probability distribution $\{p_i\}$ [1], that in special case $k = 1$ is ordinary exponential entropy (Shannon index) [5], and in special case $k = 2$ we get $E_2 = InverseHHI$, so could be named "distribution extents" of the set $\{x_1, x_2, \dots, x_n\}$.

It's important, that if distribution of revenue has density p_{Rev} then $\{p_i\}$ are normalized quantiles of empirical cumulative distribution function \mathcal{P}_{Rev} : $p_i = x_i / \sum_{i=1}^n x_i$, $x_i \approx \mathcal{P}_{Rev}^{-1}(1 - \frac{i-1/2}{n})$, rather than p_{Rev} .

4. Properties of Distribution Extents (DE)

For an arbitrary k the properties (1), (2) take form:

$$1 \leq E_k \leq n;$$

$$E_k = 1, \text{ when } x_1 = A, x_i = 0, i = 2, \dots, n \quad (4a)$$

$$E_k = n, \text{ when } x_1 = x_2 = \dots = x_n \quad (4b)$$

Let $x_i = A, i = 1, 2, \dots, n_{Large}$

$x_i = a, i = n_{Large} + 1, \dots, n$; where $a \ll A$

$$\text{Then } E_k \approx n_{Large} \left(1 + O\left(\frac{n_{Small} a}{n_{Large} A}\right) \right) \quad (5)$$

If $0 < k < m$, then

$$\frac{1}{\max(p_i)} = E_{+\infty} \leq E_m \leq E_k \leq E_0 = n \quad (6)$$

Property (5) shows that as in special case $k = 2$, if the set $\{x_i\}$ has n_{Large} large values and n_{Small} small values, then in reasonable assumptions *InverseHHI* is a little bit larger than n_{Large} .

5. Relationship between Distribution Extents and Rényi entropy

The Rényi entropy [6, 7] of order k , where $k \geq 0, k \neq 1$ is defined as

$$H_k = \frac{1}{1-k} \log(\sum p_i^k) \quad (7)$$

So

$$E_k = \exp(H_k) \quad (8)$$

6. Examples.

Example 1. DE for lognormal distribution.

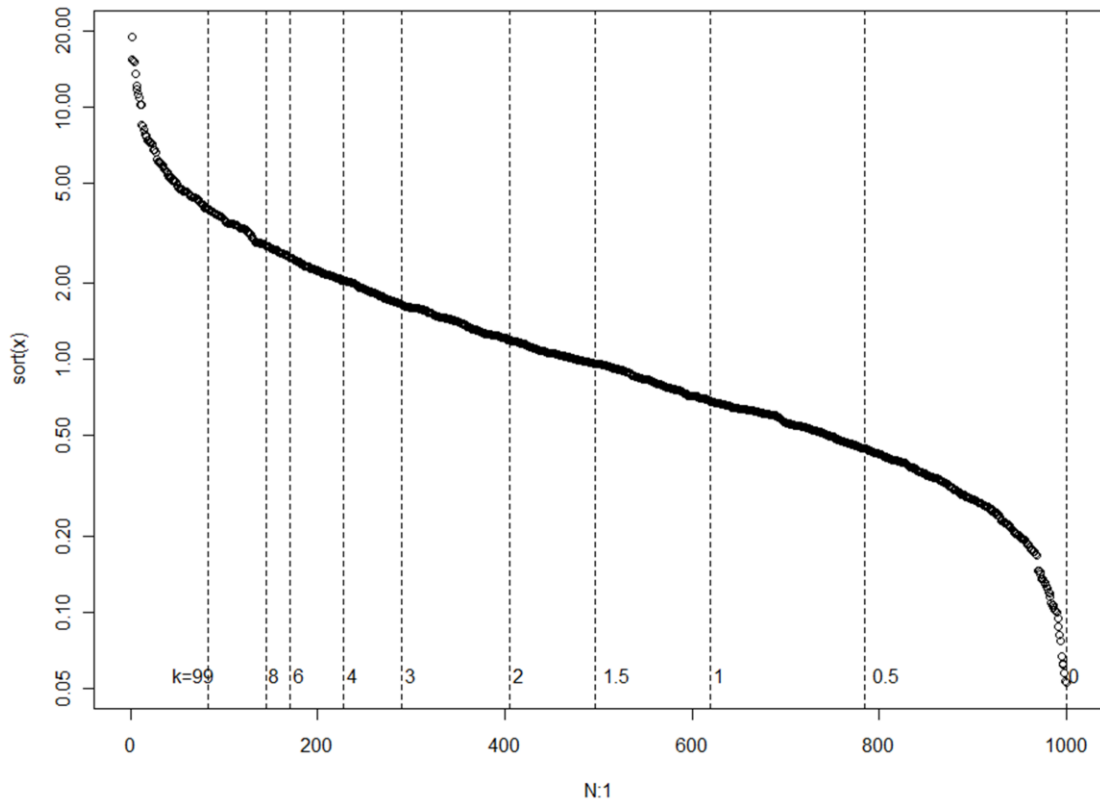


Figure 1. Positions of GEE E_k for $k \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \dots, 99\}$ on reverse quantile plot of lognormal distributed x_i , $i = 1, \dots, 1000$

In this example we created sample of 1000 lognormal distributed values x_i , sorted them in decreasing order, plotted vs. index i and marked cut-off points for values $k \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \dots, 99\}$. For given k “large x ” are left of corresponding vertical reference line.

Example 2. List of large countries (by population and GDP) for different choice of parameter k .

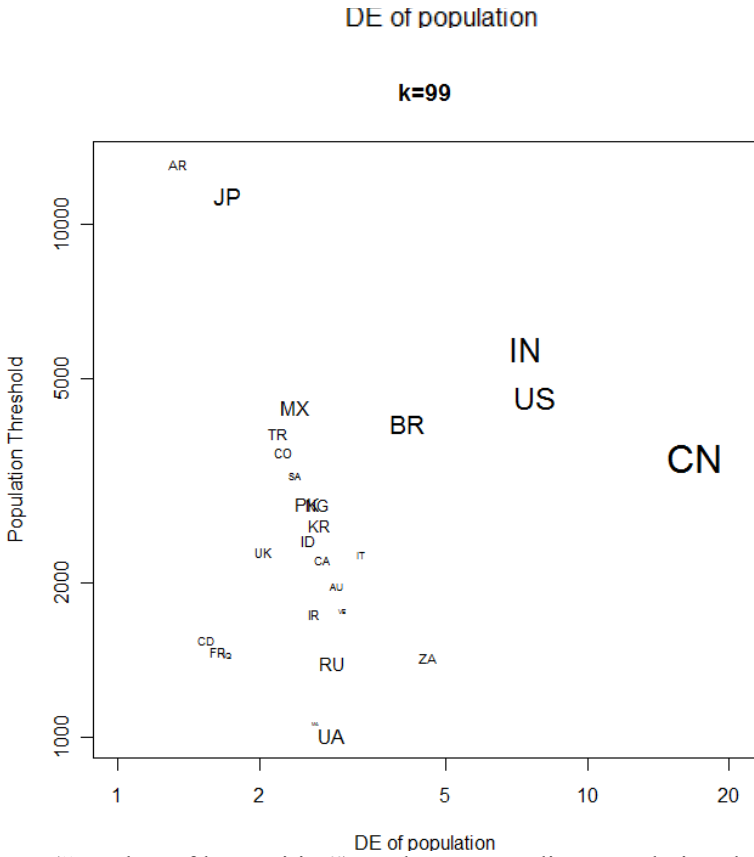
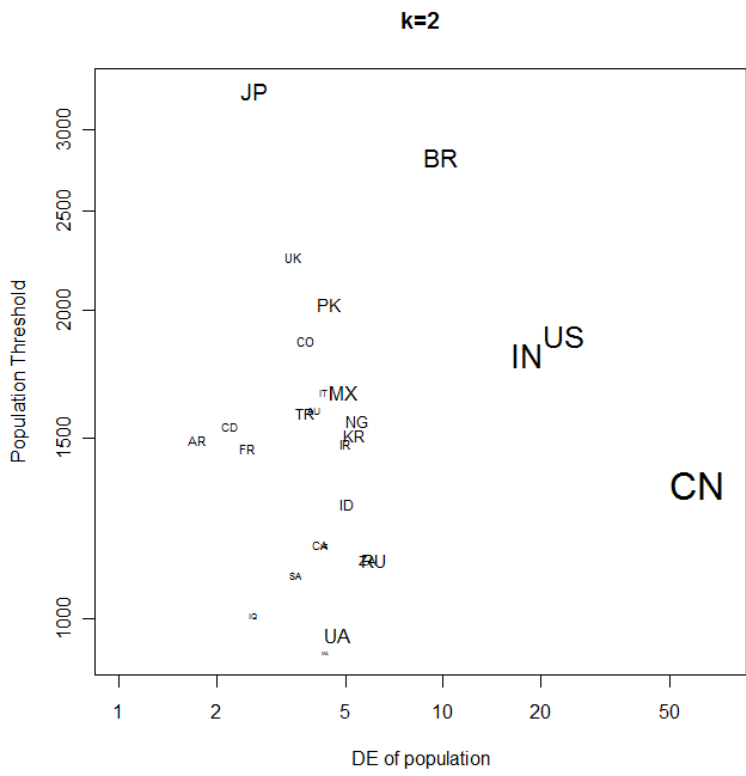


Figure 3. DE_k (“number of large cities”) and corresponding population thresholds for $k=2$ and $k=99$

We see that in both cases plots have common characteristics: China (Ch) has many large cities with smaller population in each of them, Japan (Jp) has very high concentration of population in a few huge cities with, US, Indonesia and Brasilia are somewhere in the middle.

Example 4. Usage DE in web advertising network.

We intensively used DE to analyze the Microsoft web advertising network of hundreds millions users, advertisers, and publishers that have hierarchical structure and are linked by page views, ad impressions, clicks, conversions and other characteristics that could be approximated as tripartite weighted graph with five types of edge weights. The main purpose of the analysis was community discovery for click fraud (collusion) detection. We can not describe the algorithm in use in details, because it makes “job” of fraudsters easier, so we provide here only general description of usage DE in the algorithms.

A. DE as a cut off parameters

DE as cut off parameters to present the network as a graph to analyze and visualize the network and to use graph theory methods for click fraud detection.

Parameter k of DE can be used as adjustment parameter in modeling. The picture below shows a fragment of the network graph with one type of edges, where threshold for the edges weights were chosen via DE with a high value of parameter k – keeping only edges connecting nodes with their “largest neighbors”.

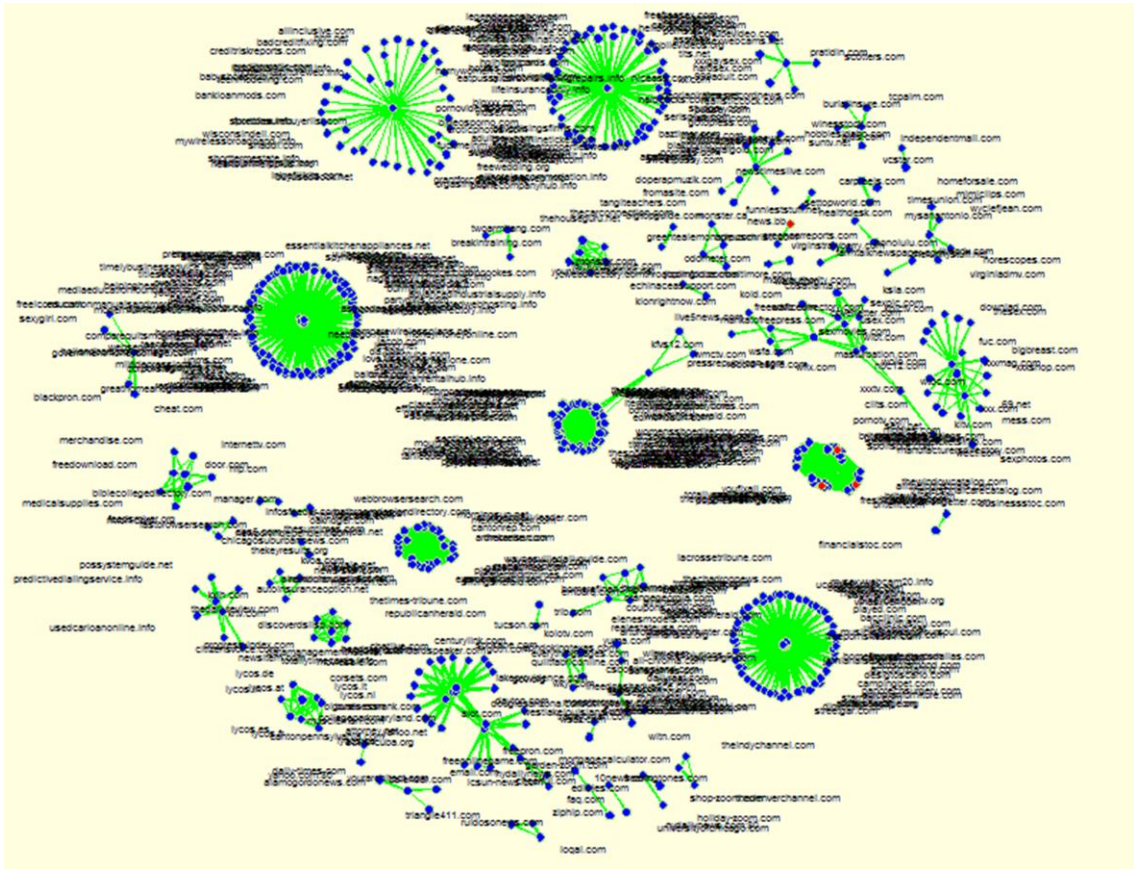


Figure 4. A fragment of the network graph with one type of edges, where threshold for the edges weights were chosen via DE with a high value of parameter k

1) *DE as an independent variable in predictive modeling.*

If we have to create a predictive model for countries, then we can use DE from example 3 above as predictive variables. Parameter k of DE can be used as adjustment parameter in modeling in this case too. The optimal value for this parameter can be found in process of cross-validation by maximization of area under the lift curve, or another objective function.

The following example illustrates usage of DE in click fraud modeling.

extent.NC	extent.F	trend/week	CV	main.DURL
2.271	12.239	-0.0403	0.3451	cooperatefinance.com
4.546	6.977	-0.0114	0.6076	topairplay.com
2.349	4.408	0.0113	1.5730	goodsearch.com
3.604	9.758	0.0103	0.7084	trainingtopic.com
3.082	5.096	-0.1361	0.3491	holidaysf.com
3.460	5.249	-0.0144	0.6814	companypetroleum.com
2.987	4.291	-0.1623	0.5622	idgeuk.com
13.156	13.989	0.0044	0.0204	facebook.com
1.112	6.269	0.0001	1.1995	autofinancecompanies-au.com

Figure 5. A fragment of a table with four predictive independent variables describing time series for a set of domains.

Variables *extent.NC* and *extent.F* are time independent DE giving number of large values in time series of some other time dependent variables NC and F for a specified time interval. The plot of a time series, where variable NC corresponds to size of bubbles, and variable F corresponds to the vertical axis, looks as following:

- [4] http://en.wikipedia.org/wiki/Simpson_index
- [5] http://en.wikipedia.org/wiki/Shannon_index
- [6] http://en.wikipedia.org/wiki/Rényi_entropy
- [7] Ziad Rached, Fady Alajaji, and L. Lorne Campbell Rényi's Divergence and Entropy Rates for Finite Alphabet Markov Sources,
<http://www.mast.queensu.ca/~fady/Journal/it01-renyi.pdf>
- [8] World Urbanization Prospects: The 2009 Revision Population Database
http://esa.un.org/unpd/wup/unup/index_panel2.html