Applied Mathematics and Computation 337 (2018) 63–75

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

An algorithm to compute data diversity index in spatial networks☆

Taras Agryzkov, Leandro Tortosa, Jose F. Vicent*

Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Alicante, Campus de San Vicente, Ap. Correos 99, Alicante E–03080, Spain

ARTICLE INFO

Keywords:
Diversity index
Spatial networks
Urban networks
Spatial statistics
Gini–Simpson index

ABSTRACT

Diversity is an important measure that according to the context, can describe different concepts of general interest: competition, evolutionary process, immigration, emigration and production among others. It has been extensively studied in different areas, as ecology, political science, economy, sociology and others. The quality of spatial context of the city can be gauged through this measure. The spatial context with its corresponding dataset can be modelled using spatial networks. Consequently, this allows us to study the diversity of data present in this specific type of networks. In this paper we propose an algorithm to measure diversity in spatial networks based on the topology and the data associated to the network. In the experiments developed with networks of different sizes, it is observed that the proposed index is independent of the size of the network, but depends on its topology.

1. Introduction

The concept of diversity has been extensively studied in different areas. For example, in ecology, diversity characterizes communities of species and ecosystems [1,2]. In political science, diversity is used to measure the effective number of political parties within parliament [3,4]. Economists use diversity to measure the effective number of firms in an industry [5]. In sociology, the ethnic diversity is studied [6]. In complex network science, the diversity of node degree is used to measure an entropy of the network itself [7,8]. In urban studies the diversity of urban facilities indicates the quality of urban environment [9,10]. Consequently, the concept of diversity has been defined in different ways, and several different indices have been developed to express it. According to the type of sample and the application context, we can highlight the following measures of diversity: Simpson concentration index [11], Shannon index [12], Rényi entropy [13] and the set of Alfa-Beta-Gama diversities [14–16]. All of these diversities depend on the contextual characteristics of a sample, data richness and data evenness [17–19]. In this sense, the richness is a measure of the total number of data types in a sample, while the evenness expresses how evenly the individual data in a sample are distributed over the different data types.

There are several areas of application of the concept of diversity very interesting to deserve. For instance, it constitutes a potential tool to promote the collective cooperation, as we can see in recent works [20,21]. In [20] the authors study the evolution of cooperation within the framework of spatial Prisoner's dilemma game with two types of players. The main

☆ Partially supported by the Spanish Government, Ministerio de Economía y Competividad, grant number TIN2017-84821-P.
* Corresponding author.
E-mail addresses: taras.agryzkov@ua.es (T. Agryzkov), tortosa@ua.es (L. Tortosa), jvicent@ua.es, jvicent@dccia.ua (J.F. Vicent).

https://doi.org/10.1016/j.amc.2018.04.068
0096-3003/

contribution of [21] is to introduce the reputation property into the interdependent lattices to further investigate the spread of cooperation within complex real-world systems. Another important application of the diversity is to identify influential spreaders in complex networks, which is conducive to deeply understanding the role of nodes in the information diffusion and epidemic spreading among a population (see [22]).

The diversity in complex networks has been studied only from the perspective of its topology. Different heterogeneity measures have been proposed in the literature to study the diversity in node degree or in the own structure of the network. But some networks, such as cities, could not be understand only from its topology. The city is a complex system [23,24], that organizes goods and different sets of features in the service of its citizens. An appropriate analysis of these elements allows us to elaborate an idea about the quality of an urban environment. Many urban studies indicate that the quality of urban space is closely related to the diversity of services located around it. A greater mix of uses and services endows the area with greater commercial activity, greater opportunities for the local business and drives a better interaction between people [9].

We have noticed that the study of data diversity in the context of the city has not been extensively studied from the perspective of spatial networks [25,26], or more specifically, urban networks [27], which represent the topology of the urban plots. The most common approach to the study of data diversity in urban environments have been made using continuous samples of data. This methodology involves the division of the city area into flat regions of a certain size or grid with the parametric refinement [10]; the issue which makes the results depend on the size and relative position of the grid. These geometric dependencies, together with the extreme sensitivity of the statistical measures of diversity, can invalidate the correct interpretation of the real data diversity in the urban context. In addition, it must be taken into account that the data in the city do not appear on a continuous plane, but are constrained to a geometry of streets. Therefore, we consider that a better approach to the issue of data diversity in the urban environment would be using the spatial network model, which allows both the modelling of the topological pattern of urban streets and the georeferencing of data present in the streets of the city.

In this paper we propose a diversity index for the data associated to a spatial network. This measure is characterized by taking into account both the local data of the node and the data associated to the neighbour nodes, following paths in the network at a specified depth.

The paper is divided into the following sections. In the first section we discuss commonly used statistical measures of diversity. In the second section, we present and justify the theoretical model to measure the data diversity within the context of spatial networks. In the third section we discuss several examples where we apply the proposed model to some networks at different scales; finally, we present some conclusions.

2. Statistical models of diversity

The basic idea of a diversity index is to obtain a quantitative measure of variability that can be used to compare different samples composed of discrete components [18]. Diversity index is a characteristic value for diversity itself, which in turn can be interpreted as the effective number of types or equally abundant types in the sample [28,29]. Diversity depends on the data richness and data evenness. Following this definition, the richness represents the number of data per sample and the evenness indicates relative abundances of the various data in a sample. The evenness increases as data are more evenly distributed in a sample such that maximum evenness is obtained when all the data are equally abundant. The value of diversity increases as the number of data per sample increases and as the abundances of data within a sample become more even [19].

Many indices for measuring data diversity have been proposed. Nearly all non-parametric diversity indices used in the sciences are monotonic functions based on the relative abundance of species $a_i = n_i/N$, with n_i referring to the abundance of the ith data in the sample, and $N = \sum_{i=1}^{R} n_i$, with R (richness) representing the total number of elements in the sample. These sets of functions include Shannon entropy, all Simpson and Gini–Simpson measures [30,31], all Rényi entropies, all *Tsallis* entropies [32], and many others. All such measures we can obtain from the reciprocal form of the weighted power mean or Hölder mean, where the proportional abundances a_i are pounded by itself. This generalized notation for diversity has been proposed by Hill [1], and it is also known in literature as *Hill numbers* or true diversity of different orders q:

$$H_q = \left(\sum_{i=1}^{R} a_i^q \right)^{\frac{1}{1-q}}. \tag{1}$$

The parameter q controls the sensitivity of the result diversity H_q to the most and least abundant elements within the sample. Thus, with $q = 0$ the diversity H_0 is completely insensitive to the frequency of elements and comes to represent the richness R. When $q < 1$ we obtain the diversities that favour disproportionately the less frequent elements in the sample and when $q > 2$ the resulting diversities favour too much the most frequent elements of the sample [33,34]. The most widely used diversities in many areas of science are H_1 and H_2, that are the limit of the exponential of Shannon entropy,

$$H_1 = \exp\left(- \sum_{i=1}^{R} a_i \ln a_i \right),$$