Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Economics and Finance 3 (2012) 62 – 67

Procedia
Economics and Finance

www.elsevier.com/locate/procedia

Emerging Market Queries in Finance and Business

Short-medium term tourist services demand forecasting with Rough Set Theory

Emilio Celotto[a,*], Andrea Ellero[a], Paola Ferretti[b]

[a]*Department of Management, Ca' Foscari University of Venice, Italy*
[b]*Department of Economics and SSE, Ca' Foscari University of Venice, Italy*

Abstract

The need to understand more in depth tourism demand trends and the aim to provide the tourist operators and the policy makers with innovative predicting tools are the key points of our research. In tourism literature predicting tourist demand has become a flourishing theme of research at a macroeconomic level, while the study is still lacking at a microeconomic level. Our attention is focused on analyzing Italian tourists' behaviours on the basis of statistical surveys on households, life conditions, incomes, consumptions, travels and holidays. Data analysis is performed by means of Rough Sets Theory, a Data Mining technique which, unlike more traditional time-series and econometric models, can easily manage categorical variables. Data were provided by GfK Eurisko and concern social, cultural and behavioural trends in Italy, collected by means of a psychographic survey. Some interesting relations between consumer behaviours and corresponding tourism consumption choices are obtained in terms of decision rules.

Keywords: Rough Sets, Tourism, Forecasting, Data Mining

1. Introduction

Traditional non-causal time-series and causal econometric models are consolidated tools for macroeconomic forecasts and allow to predict tourist flows arrivals, room nights or turnover expenditure, revenues, etc. on the basis of macroeconomic variables like GDP, currencies exchange rates or demographic variables. Their

* Corresponding author.
E-mail: ellero@unive.it

doi:10.1016/S2212-5671(12)00121-9

informative content is limited since they are weak in predicting tourist demand at a microeconomic level, on the basis of data like consumptions, travels and vacation or household budgets. Our main research goal was to identify, in a database provided by GfK Eurisko, a set of attributes concerning non-tourism consumption behaviors which are strictly related to the Italian tourist demand. More precisely, we were looking for non-tourism products or services whose consumption could possibly anticipate tourists' behavior, as "sentry variables" to be used by the tourism sector stakeholders to forecast short-term to medium-term tourism demand trends see e.g. Kulendran and Witt, 2003. We mined therefore the dataset looking for robust ties among variables expressed in term of decision rules. For the mining process we used Rough Sets Theory see e.g. Pawlak, 1991 which has a mix of remarkable features: no assumptions about data are required; no need to define *a priori* functions or equations; categorical data are accepted; relationships between data are explained by simple inductive rules. The paper is organized as follows. Section 2 is devoted to the introduction of the main concepts of Rough Sets Theory. Section 3 analyzes the data set and depicts the main characteristics of the used software package. Section 4 presents the main rules discovered by the methodology. We address validation of the rules in Section 5 while Section 6 ends the paper with some conclusive remarks and hints for future research.

2. Rough Sets and applications to tourism

Rough Sets Theory has been presented as *a new mathematical tool for imperfect data analysis* in Pawlak 1991. It concerns approximated knowledge of data by means of two alternative approaches: lower/upper approximations of a data set and decision rules. A decision rule of the form "conditions ⇒ decision" can be mined from the dataset and two indexes, *coverage* and *certainty*, give a measure of the goodness of the rule Stefanowski, 1998. The model adopted in our research is based on a set of single condition attribute rules, grouped by subject. It is modular in the sense that it can be enriched with new rules as soon as new data are available. The choice of a single condition is due both to our aim to identify possible products/goods which consumption can be interpreted as sentry variables and to the need of easy to understand/communicate rules. The indiscernibility relation for which all objects characterized by the same information are *indiscernible* is the key mathematical concept of the theory. A pair $S=(U,A)$ composed by a finite nonempty universe U of elements $\{x_1,x_2,\ldots,x_N\}$ and a finite set A of attributes $\{a_1,a_2,\ldots,a_k\}$ is called an information system. For each $B\subseteq A$, the universe U is split into a family of equivalence classes called *elementary sets* through the indiscernibility relation stating that two objects x_i and x_j in U cannot be distinguished with reference to the set of attributes in B. When the elements of a *concept* X, i.e. a set $X\subseteq U$, are all in elementary sets, then its objects can be distinguished in terms of the available attributes; otherwise, X is *roughly* defined. The information system can be denoted by $S=(U,C,D)$ if the attributes are divided into two disjoint sets C and D, of condition and decision attributes, respectively. The *decision rule induced by object x in* S is the a sequence $c_1(x),c_2(x),\ldots,c_n(x),d_1(x),d_2(x),\ldots,d_m(x)$, in short $C\rightarrow_x D$, when n condition attributes and m decision attributes are identified. The *strength* $\sigma_x(C,D)$ *of the decision rule* $C\rightarrow_x D$ is represented by the ratio between the support of the decision rule and the cardinality of U. The *certainty factor* $cer_x(C,D)$ and the *coverage factor* $\mathrm{cov}_x(C,D)$ of the decision rule are defined as the conditional probability $\pi_x(D\,|\,C)$ that $y\in D$ conditionally to the assumption that $y\in C$ and, respectively the conditional probability $\pi_x(D\,|\,C)$ that $y\in D$ conditionally to the assumption that $y\in D$. The certainty factor can be defined as the "rule probability" and measures how strong the "conclusions from the data" are.