Using Web scrapping techniques for price statistics - the Romanian experience

Dr. Bogdan OANCEA

Director, Dept. of Innovative Tools in Official Statistics, INS, Romania

Abstract. Internet has been widely recognized as a new data source that can be used to compile new statistics or the enhance the traditional ones in several fields of official statistics. Considering that online commerce has a growing share in the overall household's consumption, price statistics is one of the areas of official statistics that can have important benefits from this new data source. There have been several projects around European countries, exploring the potential of Web scrapping techniques to enhance the production of the classical consumer price index. In this paper we will describe the experience of NSI Romania regarding the collection of prices from Internet and compiling a consumer price index. The aim of our pilot project was to investigate whether the Web scrapping method of data collection for prices can be introduced in the production of official statistics in the near future and what are the methodological challenges that we have to deal with. We developed a chain of tools that automates the whole process, starting with data collection, transforming the semistructured data into structured data, going to a data validation procedure and finally to a computation procedure that outputs a price index. We started from the traditional methodology used for CPI but we added some new features such as a clustering technique and a distance-based method for matching similar products to take advantage of the specificity of the web-scrapping collection method. The whole process was represented in terms of GSBPM and the quality of estimates has been investigated.