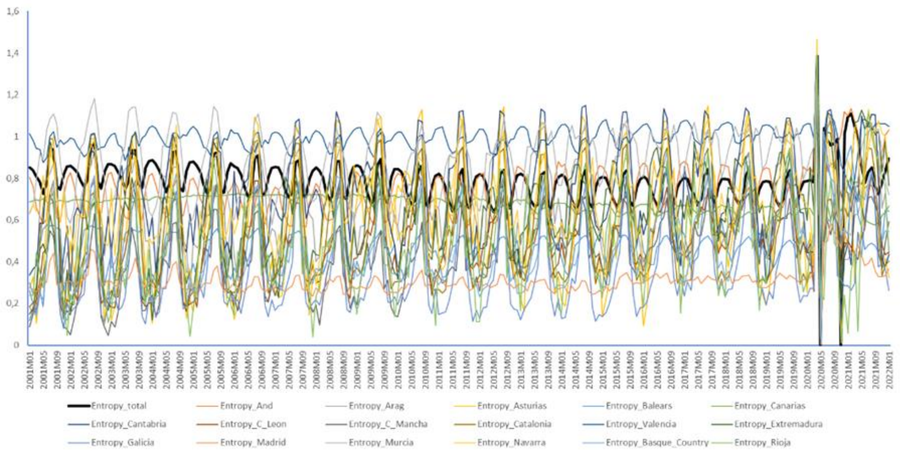
# Multichoice Entropy Clustering for Time Series and Seasonality




**Miguel Ángel Ruiz Reina**
**ruizreina@uma.es**
**Department of Economic Theory and Economic History (THE)**
**Universidad de Málaga**

UNIVERSIDAD DE MÁLAGA

# Highlights

- **Dimensional reduction** for time series based on **Information Theory**.
- New automatic **cluster method classifies** information based on Shannon entropy.
- The **internal verification** guarantees that the clusters present space-time similarities.
- The **unsupervised cluster** method automatically adjusts for the number of spatial and temporal observations under study.
- Our **empirical approach** allows us to measure accommodation decisions for **multichoice tourists** who visit Spain and decide their **accommodations**.
- The results show that the clusters are **dynamic and fit the analysed data set**.

- **Information Theory** to recognise behaviour patterns in the **decision-making.**
- **Statistical Learning** models usually combine **Statistics and Machine Learning** in contexts of uncertainty, relaxing the limiting assumptions of Statistics to large volumes of data (although it is possible with smaller data dimensions).
- **Data science (using Statistical Learning) is the ability to turn data into value, for example by iterating through this process:**
    1. Understand the business problem
    2. Understand the data
    3. Prepare the data
    4. Build a model
    5. Evaluate
    6. Deploy

**“It doesn't matter what tools you use for this.** But you should be able to go all the way from business understanding to deployment without too much handholding and with a reproducible scheme that's reliable - if needed on a large scale”. (**Tobias Zwingmann. Senior Data Science).**

## Business issues:

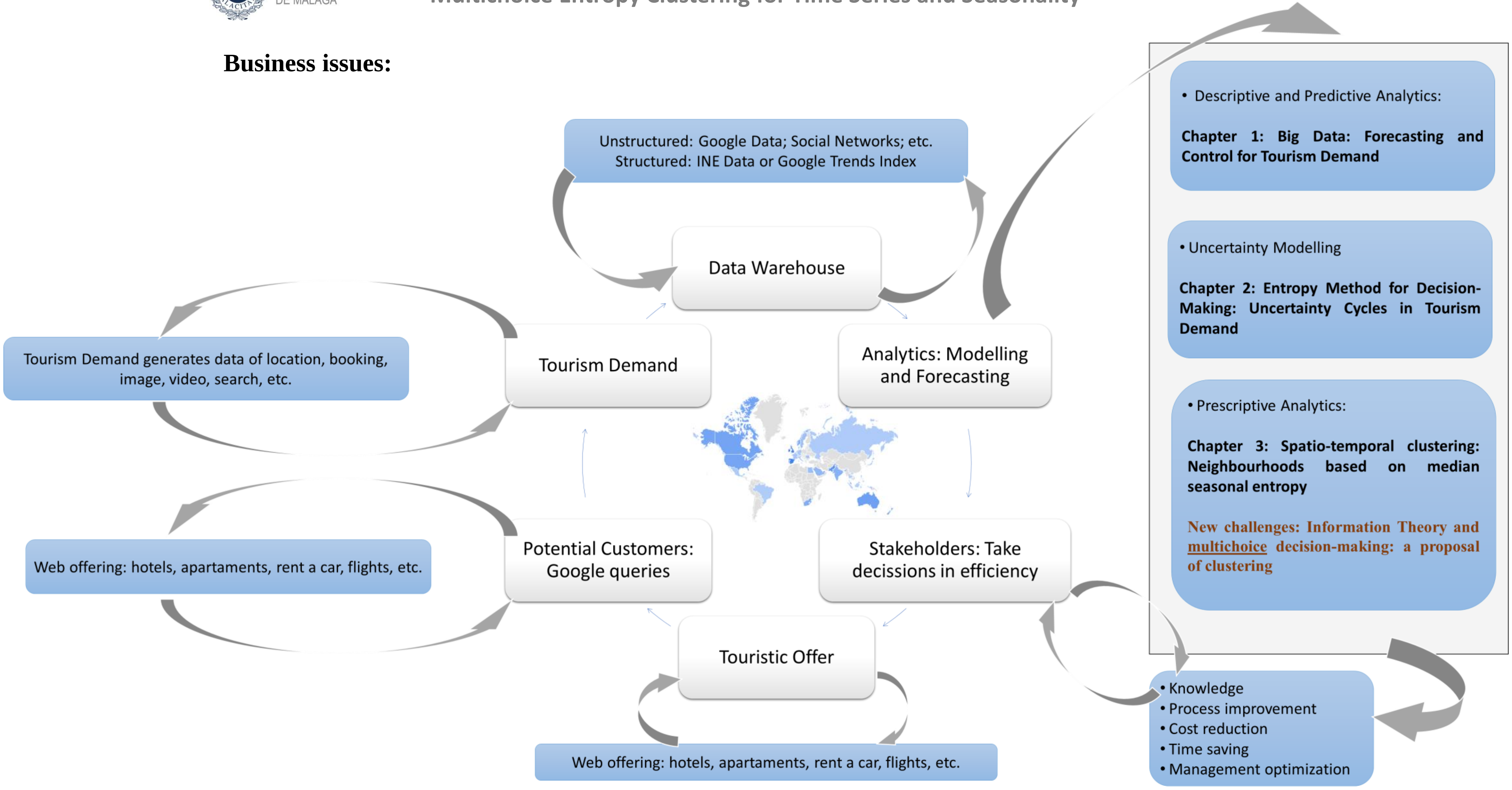



Fig. 1. **The life cycle of the data in the Tourism Demand**. Own elaboration.

## Tourism industry and relevance in modern economies: the case of Spain

- Spain: finds itself in a **political, social, economic, cultural and demographic environment** conducive to developing the world's fastest-growing industry in the service sector.

- **INE 2019**:
    i. 12.4% to Spanish GDP (including 6.4% directly and 6% indirectly).
    ii. 12.8% of the workforce in the labour market.

- Spain in 2019 represents the **second largest number** of people in the world (82.7 million) of the arrival of people [France (89.4 million); USA (79.6 million); China (62.9 million), Italy (62.1 million); Turkey (45.7 million] **(UNWTO, 2021).**

- **World Economic Forum** revealed that Spain is the most prepared country for the tourism industry with an index of **5.4 in 2019** ahead of economically more developed countries in other industries such as France, Germany, Japan, United States, United Kingdom, Australia, Italy, Canada or Switzerland (Uppink-Calderwood, 2019).

"… Spain is one of the world's top tourism destinations and extremely depends on tourism" (Camacho & Pacce, 2017)

**Main ideas:**

- **Clustering** metric methods are crucial for many **real-world applications**, and distance metrics provide learning models with better performance than is generally achieved (Qin et al., 2015).
- **Unsupervised learning** automatically generates categorisation problems into new categories for the same attributes as each other in machine learning operations (Chen & Yin, 2018).
- **Clustering techniques are variants for time series reducing dimensionality**; they constitute a wide area of research (Aghabozorgi et al., 2015).

**Short notes of literature review**

- **Clustering interesting for disciplines**: Economics, Scientific, Computational, Health or Natural.
- Since the 1960s, three main questions to solve: **How many clusters are there? What is the best clustering algorithm? What can we do with the outliers?**
- **Clustering and classification techniques play a central role** in exploiting those, for instance, in **Big Data environment for machine learning techniques** (Luna-Romera et al., 2018)

## Some works using Clustering- Entropy:

- **Entropy-based uncertainty** for space-time analysis has been used in **geosciences** (Leibovici et al., 2020).
- **Categorical variables** of interest on a dataset linked to climate change and evolution of land cover types in **Nordic areas** (Leibovici & Claramunt, 2019).
- For non-Gaussian spatial random fields is explored for spatial field reconstruction in environmental and sensor network monitoring (Peters et al., 2021).
- **Analyse traffic predictions** in Smart urban areas (S. Zhou et al., 2021).

**Entropy and decision-making in Tourism** (Ruiz-Reina, 2021)

- "Entropy Method for Decision-Making: Uncertainty Cycles in Tourism Demand". (Journal: Entropy)
- "Spatio-temporal clustering: Neighbourhoods based on median seasonal entropy". (Journal: Spatial Statistics)

**3. Methodology**

**3.1. A general scheme: HDTS to RDTS (multichoice)**

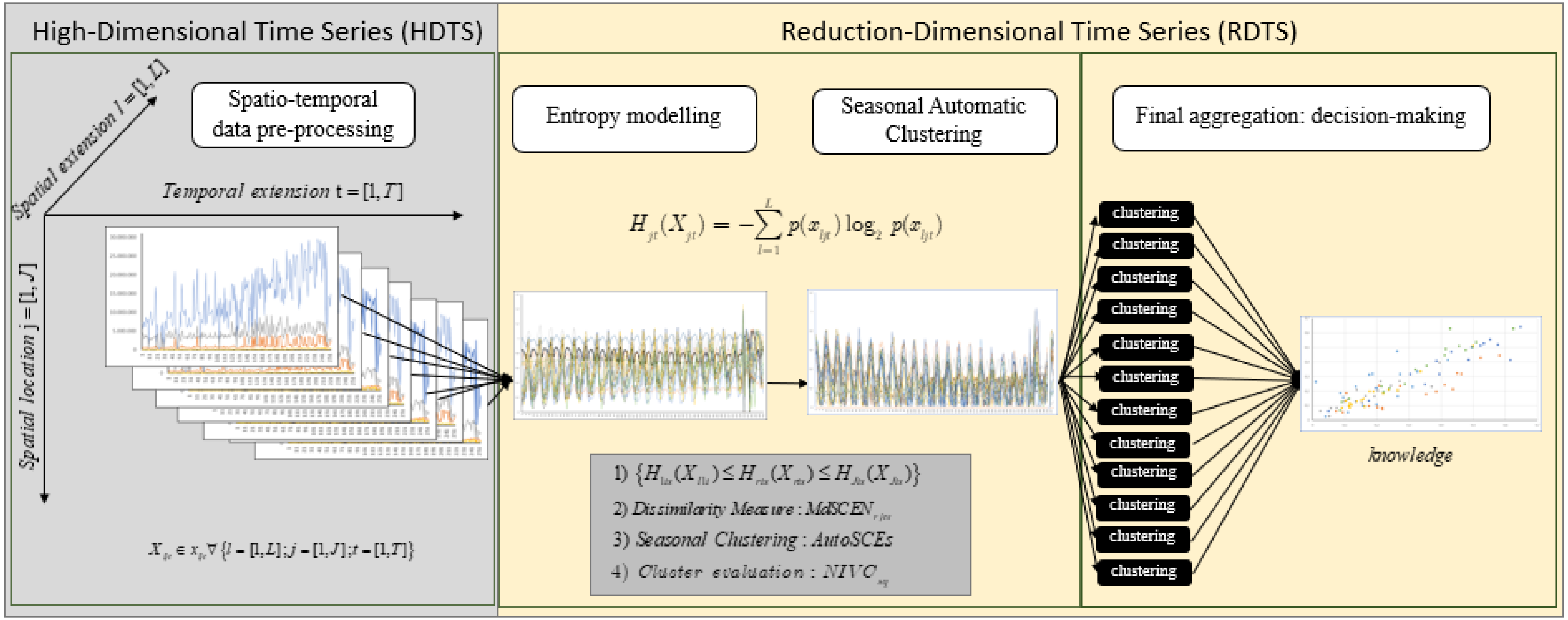



**Fig. 2.** AutoSCEs: phases for High-Dimensional Time Series (HDTS) to Reduction-Dimensional Time Series (RDTS). Own elaboration.

3. Methodology

## 3.2. A general scheme: HDTS to RDTS (multichoice)

**HDTS:** Formally the set of time series analyses

$$x_{ljt} \in X_{ljt} \forall \left\langle \underbrace{l = 1,2,\ldots,L}_{\text{what type of accommodation ?}} ; \underbrace{j = 1,2,\ldots,J}_{where\,?} ; \underbrace{t = 1,2,\ldots,T}_{when\,?} \right\rangle$$

**RDTS:** Where *l* represents the options of the variables *x* in the space *j* for a temporal analysis period of each *t* in the data set (temporal space of *t* is divided into seasonal flows *s*). Entropy and reference time series (Shannon, 1948):

$$H_{jt}(X_{jt}) = -\sum_{l=1}^{L} p(x_{ljt}) \log_2 p(x_{ljt})$$

$$e.g.:$$

$$l = 1\ (hotels), l = 2\ (apartments), l = 3\ (campsites)\ and\ l = 4\ (rural\ apartments)$$

$$j = 1\ (Andalusia), j = 2\ (Aragon),\ \ldots\ , j = 17\ (Valencian\ Community)$$

$$t = 2001.1,\ \ldots\ , 2022.1$$

$$where\ (x_{ljt}) = \left| x_{ljt} \right| \Big/ \left( \sum_{l=1}^{L} x_{ljt} \right) \leq 1\ and\ \sum_{l=1}^{L} (x_{ljt}) = 1$$

$$The\ centroid\ of\ the\ clustering\ is\ the\ reference\ series\ H_r(x_{rt})$$

$$\min H_{jt}(X_{jt}) \leq H_r(x_{rt}) \leq \max H_{jt}(X_{jt})$$

3. Methodology

## 3.2. A general scheme: HDTS to RDTS (multichoice)

**Part 1** (starting reduction time series):

$$\overrightarrow{H_{jt}(X_{jt})} \ \ and \ \ \overrightarrow{H_{rt}(x_{rt})}$$
$$for \ \ seasonality$$
$$\overrightarrow{H_{jts}(X_{jts})} \ \ and \ \ \overrightarrow{H_{rts}(x_{rts})}$$

**Part 2** (measuring distance of reference series):

$$MdSCEN_{r\,jts} = Median\left(\sqrt{H_{rts}(x_{rts}) - H_{jts}(x_{jts})}^{\,2}\right)$$

**Part 3** (deciding automatically number of clusters *Auto Seasonal Clustering Entropy - AutoSCEs*):

$$N = \sum_{i=0}^{AutoSCEs} \binom{AutoSCEs-1}{i} = (1+1)^{AutoSCEs-1}$$
$$where:$$
$$N \ \ represents \ \ numbers \ \ of \ \ observation \ \ in \ \ j \ \ locations$$
$$i \ \ contains \ \ the \ \ number \ \ of \ \ samples$$
$$Finally, \ \ applying \ \ logarithms \ \ shows \ \ that \ \ the \ \ optimal \ \ number \ \ of \ \ clusters \ \ is \ \ expressed \ \ by:$$
$$AutoSCEs = 1 + \log_2(N)$$

**Part 4** (Cluster evaluation):

- The **maximum distance intracluster** is defined as the breadth of a cluster *i* $BC_i = \dfrac{\max H_{jts}(X_{jts}) - \min H_{jts}(X_{jts})}{AutoSCEs}$
- **Neighbourhood's Internal Verification Coefficient**:

$$NIVC_{sq} = \frac{\max[d(MdSCEN^{u}_{srjt}, MdSCEN^{l}_{srjt})]}{BC_i}$$

$$0 \leq NIVC_{sq} \leq 1$$

the most remarkable intra-group **similarity** will occur when $NIVC_{sq} = 0$

otherwise we will find greater **diversity** with $NIVC_{sq} = 1$

**Part 5 (**Final aggregation: decision-making):

The neighbourhood-based clusters are obtained: $cluster_q = cluster_1, cluster_2, \ldots, cluster_Q$.

**Part 6: Exploiting knowledge for each field**

4. A case of study

## 4.1. Theoretical Analysis framework for decision-making in Tourism demand

4.2. HDTS to RDTS: foreigner demand in the Spanish tourism accommodation market

4.3. RDTS: Entropy time series for decision-making

4.4. Clustering results

"Clusters and the new economics of competition": **clusters are defined as geographic concentrations of interconnected companies, specialised suppliers and consumers, and associated institutions (Porter, 1998)**

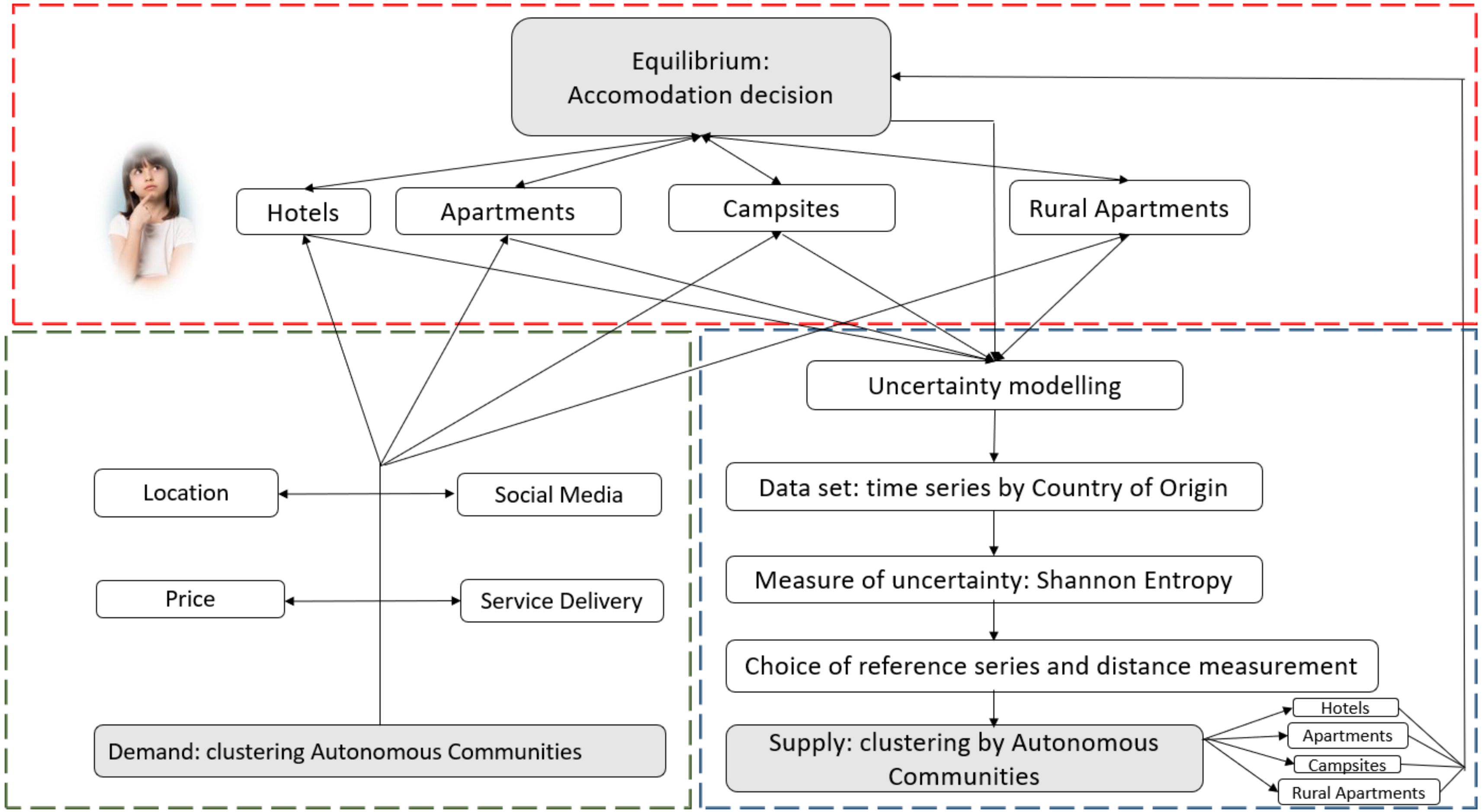



Fig. 3 Research question for Tourism demand. **Theoretical Analysis framework for Tourism demand.**

Database used comes from the **Spanish National Institute of Statistics (INE)**

**Accommodation - Multichoice data for foreigners:**

***l***= *(1= overnight stays in hotels, 2 = campsites, 3 = tourist apartments and 4= rural apartments)*

***j*** **=** *(1= Andalusia, 2= Aragon, 3= Asturias, 4= Balearic Islands, 5= Canary Islands, 6= Cantabria, 7= Castile Leon, 8= Castilla la Mancha, 9= Catalonia, 10= Valencia, 11= Extremadura, 12= Galicia, 13= Madrid, 14= Murcia, 15= Navarra, 16= Basque Country, 17= La Rioja)* and the added variable called ***"Total"*** **is the reference series *r***.

**Seasonality:** from January 2001 to January 2022 *(t=1.2, … , T) and* **s =** January, February, March, …, December

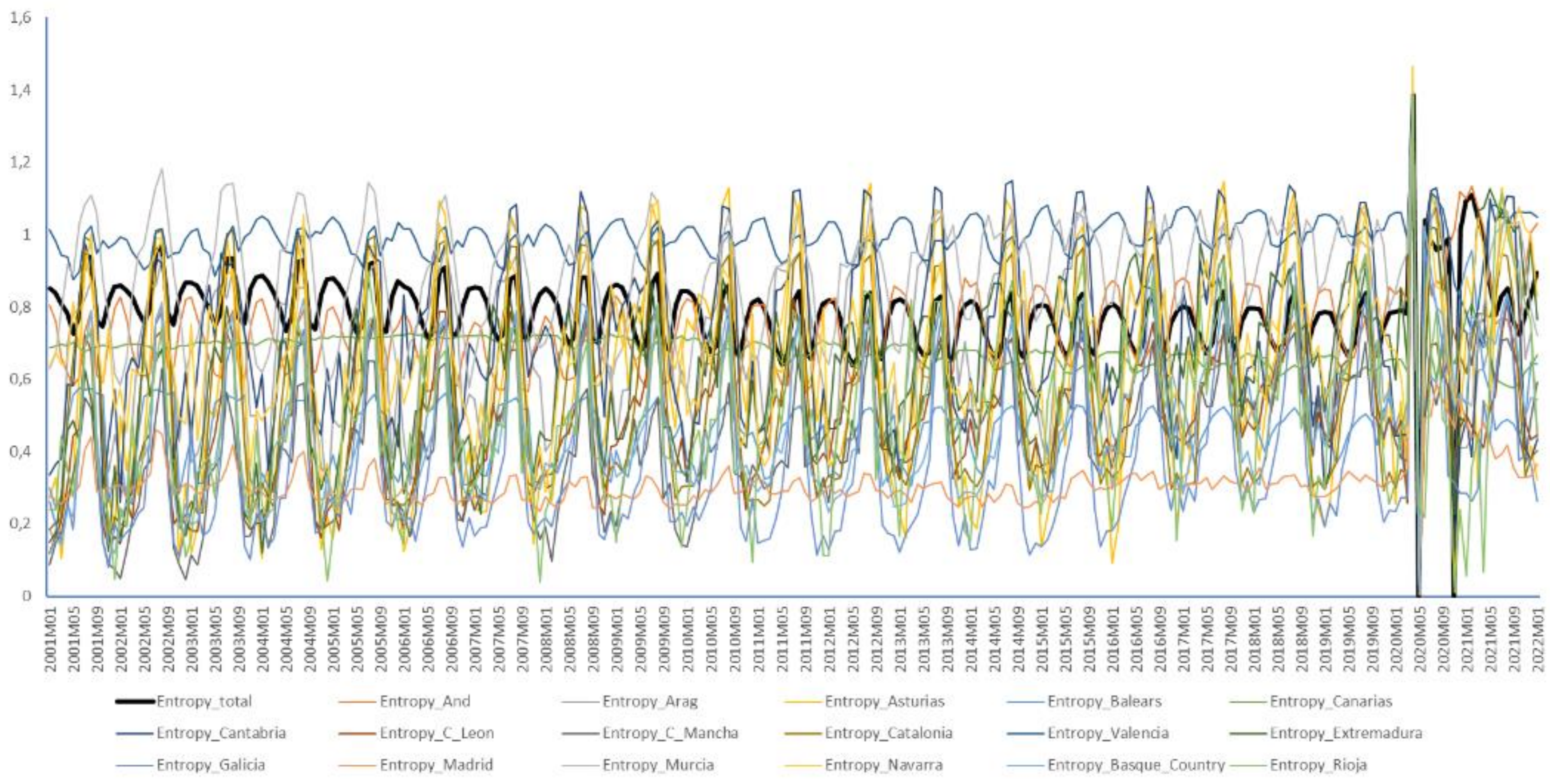



**Fig. 4.** Measurement of spatial uncertainty based on Entropy Between **Jan. 2001 to Jan. 2022**. Data from INE. Here the reference series r is called Total. Own Elaboration.

4. A case of study

4.1. Theoretical Analysis framework for decision-making in Tourism demand

4.2. HDTS to RDTS: foreigner demand in the Spanish tourism accommodation market

4.3. RDTS: Entropy time series for decision-making

## 4.4. Clustering results

# Multichoice Entropy Clustering for Time Series and Seasonality

| | January | | | | February | | | | March | | | | April | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | limits | | | | limits | | | | limits | | | | limits | | | |
| | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC |
| 1 | 0,00 | 0,12 | 1 | 0,00 | 0,00 | 0,11 | 1 | 0,00 | 0,00 | 0,10 | 2 | 0,21 | 0,00 | 0,08 | 3 | 0,18 |
| 2 | 0,12 | 0,24 | 3 | 0,54 | 0,11 | 0,23 | 3 | 0,82 | 0,10 | 0,20 | 2 | 0,70 | 0,08 | 0,15 | 3 | 0,17 |
| 3 | 0,24 | 0,35 | 2 | 0,39 | 0,23 | 0,34 | 3 | 0,63 | 0,20 | 0,30 | 5 | 0,71 | 0,15 | 0,23 | 4 | 0,44 |
| 4 | 0,35 | 0,47 | 3 | 0,40 | 0,34 | 0,45 | 3 | 0,07 | 0,30 | 0,40 | 2 | 0,32 | 0,23 | 0,31 | 3 | 0,11 |
| 5 | 0,47 | 0,59 | 6 | 0,78 | 0,45 | 0,57 | 6 | 0,74 | 0,40 | 0,50 | 4 | 0,53 | 0,31 | 0,38 | 2 | 0,05 |
| 6 | 0,59 | 0,71 | 2 | 0,27 | 0,57 | 0,68 | 1 | 0,00 | 0,50 | 0,60 | 2 | 0,59 | 0,38 | 0,46 | 2 | 0,21 |
| 7 | 0,71 | 0,82 | 0 | - | 0,68 | 0,79 | 0 | - | 0,60 | 0,70 | 0 | - | 0,46 | 0,54 | 0 | - |
| | | total | 17 | | | total | 17 | | | total | 17 | | | total | 17 | |

| | May | | | | June | | | | July | | | | August | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | limits | | | | limits | | | | limits | | | | limits | | | |
| | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC |
| 1 | 0,00 | 0,07 | 3 | 0,44 | 0,00 | 0,07 | 2 | 0,79 | 0,00 | 0,08 | 0 | - | 0,00 | 0,08 | 0 | - |
| 2 | 0,07 | 0,15 | 4 | 0,86 | 0,07 | 0,15 | 5 | 0,80 | 0,08 | 0,16 | 9 | 0,83 | 0,08 | 0,17 | 10 | 0,74 |
| 3 | 0,15 | 0,22 | 5 | 0,80 | 0,15 | 0,22 | 5 | 0,62 | 0,16 | 0,24 | 5 | 0,94 | 0,17 | 0,25 | 5 | 0,32 |
| 4 | 0,22 | 0,29 | 3 | 0,37 | 0,22 | 0,29 | 4 | 0,53 | 0,24 | 0,32 | 1 | 0,00 | 0,25 | 0,34 | 1 | 0,00 |
| 5 | 0,29 | 0,36 | 0 | - | 0,29 | 0,36 | 0 | - | 0,32 | 0,40 | 1 | 0,00 | 0,34 | 0,42 | 0 | - |
| 6 | 0,36 | 0,44 | 2 | 0,33 | 0,36 | 0,44 | 1 | 0,00 | 0,40 | 0,48 | 0 | - | 0,42 | 0,50 | 0 | - |
| 7 | 0,44 | 0,51 | 0 | - | 0,44 | 0,51 | 0 | - | 0,48 | 0,56 | 1 | 0,00 | 0,50 | 0,59 | 1 | 0,00 |
| | | total | 17 | | | total | 17 | | | total | 17 | | | total | 17 | |

| | September | | | | October | | | | November | | | | December | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | limits | | | | limits | | | | limits | | | | limits | | | |
| | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC | lower | Upper | Elements | NIVC |
| 1 | 0,00 | 0,08 | 2 | 0,38 | 0,00 | 0,09 | 2 | 0,35 | 0,00 | 0,11 | 2 | 0,05 | 0,00 | 0,11 | 1 | 0,00 |
| 2 | 0,08 | 0,16 | 5 | 0,85 | 0,09 | 0,18 | 4 | 0,73 | 0,11 | 0,21 | 1 | 0,00 | 0,11 | 0,22 | 4 | 0,87 |
| 3 | 0,16 | 0,24 | 4 | 0,29 | 0,18 | 0,26 | 5 | 0,49 | 0,21 | 0,32 | 4 | 0,69 | 0,22 | 0,33 | 2 | 0,14 |
| 4 | 0,24 | 0,32 | 4 | 0,59 | 0,26 | 0,35 | 4 | 0,79 | 0,32 | 0,42 | 4 | 0,88 | 0,33 | 0,44 | 2 | 0,25 |
| 5 | 0,32 | 0,40 | 1 | 0,00 | 0,35 | 0,44 | 1 | 0,00 | 0,42 | 0,53 | 5 | 0,76 | 0,44 | 0,55 | 6 | 0,82 |
| 6 | 0,40 | 0,49 | 1 | 0,00 | 0,44 | 0,53 | 1 | 0,00 | 0,53 | 0,64 | 1 | 0,00 | 0,55 | 0,66 | 2 | 0,00 |
| 7 | 0,49 | 0,57 | 0 | - | 0,53 | 0,62 | 0 | - | 0,64 | 0,74 | 0 | - | 0,66 | 0,77 | 0 | - |
| | | total | 17 | | | total | 17 | | | total | 17 | | | total | 17 | |

Table 1. **Neighbourhood's internal verification coefficient** for the clusters found in the treatment process according to nationalities that visit Spain in the intra-annual period January to December. Values close to **zero indicate maximum similarity** between the cluster members when **close to one; they indicate maximum dissimilarity**. Own Elaboration.

## Conclusions

- This Clustering System contributes (Try) to **Thermoeconomics and Physical Statistics** with application to Economics in Information Theory applied to dynamic time series structures.
- **The need for robust, simple models for complex realities seems to be closely linked (Lisciandra, 2016).**
- In general, clustering methods **minimises the risk and provides knowledge** in decision-making.
- Clustering processing: for the **organisation of seasonal Spatio-Temporal information**.
- The **seasonal clustering problem can be solved efficiently.**
- This study **reveals some implications** for geostatistical research and the clustered tourist accommodation market based on the empirical results (Michael, 2003).
- **Previous studies revealed differences in accommodation decisions among** nationalities (20) and seasons (Ruiz-Reina, 2021). **Now, a proposal for multichoice approach (**hotels, apartments, campsites or rural apartments).
- This work contributes to the **mapping and understanding to find the supply-demand equilibrium of international tourist accommodation markets and secondary industries** (Vlachos & Bogdanovic, 2013; Adamiak, 2018).
- The empirical section results **respond to non-dynamic limitations of dynamic clustering** of the literature in the tourism sector (Yong-Jin et al., 2020).

## 5.2. Future Research Lines

Other applications:

- **Transport** (car, train or aeroplane), **destinations** (beach, mountain or city), **types of trips** (business, leisure or cultural) and a **wide range of possibilities**.
- **Finance** (real estate investment in housing or rustic land), **health** (measurement of risk uncertainty), **automobiles** (selection of types of cars to buy), **pandemics** (infected population vs non infected population), etc.
- **Spatial Statistics applications** a source of future research in any field (Gelfand, 2020)
- **Child mortality** (Morales-Otero & Núñez-Antón, 2021).
- **Big Data** (Banerjee, 2020) and Small Data **analysis**.
- **Animal movements** (Hooten et al., 2020).
- **Pandemics.**

In short, **Spatial Statistics has been placed among the first lines of research in the last 15 years**; mainly, this work represents a beginning towards new frontiers of geospatial data in discrete time, for which additional lines would be the development of clustering models in continuous data.

**"Unsupervised clustering criteria applied to time series** are less developed than traditional analyses on cross-section data" (Aghabozorgi et al., 2015)

# Thank you!

**Miguel Ángel Ruiz Reina**
**ruizreina@uma.es**
**Department of Economic Theory and Economic History (THE)**
**Universidad de Málaga**