How much is it? An innovative methodology to measure room rates across OTAs and online platforms

M. Ariel Juárez[1], M. Pedro Anzo[2], M. Arturo Constantini[3]

Introduction

Tourism has become a key driver for socio-economic progress in Mexico. The sector accounts for 8.7% of the Gross Domestic Product (GDP), overcoming industries such as the automobile, finance, and construction, and employing ten million people directly or indirectly. In 2017, Mexico became the sixth most visited country in the world with 39.3 million international tourists.

The Ministry of Tourism has a clear objective of strengthening product diversification and encouraging inclusive and sustainable growth through competitive strategies and rigorous data analysis, allowing us to understand tourist behavior and identify the current trends within the sector.

In 2001, the Ministry started to develop a new statistical design to measure hotel occupation through its own system, named "DataTur." Hotel monitoring is based upon a traditional methodology of surveys in which data is collected via telephone, online or visiting the business establishment in person.

However, the fast-paced growth of the Internet allows innovative ways to offer products, generating new business models based on e-commerce. Because of that, the online ecosystem is considered to be a valuable source of information. The goal is to exploit it by obtaining the average rate for every tourist destination, measure the extra-hotel sector and share economy, and gathering updated information about these accommodations.

The methodology requires a technological tool that should take into consideration the constant inflows of information from the Online Travel Agencies (OTAs) and Airbnb websites.

[1] Ministry of Tourism; ajuarez@sectur.gob.mx [2] Ministry of Tourism; aanzo@sectur.gob.mx [3] Ministry of Tourism; aconstantini@sectur.gob.mx

1. Background

Currently, many activities related to tourism are carried out through the Internet, which has generated a large amount of data that could be obtained and analyzed in real time. Consequently, today there is work based on the development of practical methodologies to conduct tourism research, using data from websites, aimed to generate information that contributes to traditional methodologies for the design of competitive strategies required by the sector.

Many of the existing works are related to data extraction on comments that tourists provide on websites to measure customer satisfaction and create indicators for decision making. However, few are related to the goal of this project: identify, quantify and monitor the accommodation offer of the hotel and extra-hotel sectors with a basic methodology; specifically the offer that is considered as the number of accommodations offered in the OTAs and the average weekly rate of each tourist destination.

For example, Guizzardi et. analyze the prices offered for the regions of Rome and Milan, considering a one year period, between June 2015 and June 2016. The data came from Expedia.com and was collected through the Rate Tiger provider, establishing as search parameters a person and a single night stay.

Also, J. Kepa-Gerrikagoitia et. proposed a competitive intelligence tool to carry out market studies of the Spanish hotel sector. The tool calculates the evolution of rates, estimates the demand and predicts the availability of the rooms within a period of 14 days, so they propose the development of a knowledge generation process, divided into three phases, in which the first one corresponds to the extraction of data through the use of the web crawling technique: browse through the Internet in a systematic way.

As mentioned, there are practical methodologies and studies related to the extraction of OTA's rates (from the traditional and shared economy sectors) but not about the construction of housing directories based on the extracted data. The reviewed works perform the extraction of information using web crawling on one or two specific websites, to estimate the average daily rate as an indicator of tourist destinations in Spain and Italy. However, each of the articles uses a different methodology from the one presented in this paper, using more than one source of information and analyzing 70 tourist destinations, covering both the hotel and extra-hotel sectors.

2. Methodology

2.1. Summary

In this methodology, the accommodation offer is related to the timely obtaining of the directory of traditional and non-traditional accommodations, as well as the average weekly rate offered online, by the accommodations in each tourist destination. The scope of this methodology is determined by the availability of the information provided by the data sources used (online travel agencies OTAs and Metasearch engines). The methodology is applied with a technological tool that extracts and processes data to build and store housing directories in established times, allowing the user to arrive at conclusions about historical data.

Collecting information: Obtaining information is done with a software tool and considers three main processes:

1) **Web search and extraction:** Search of the websites of interest and extract the available information from the accommodations.

2) **Data processing:** Execute text processing techniques to obtain only the variables of interest.

3) **Storage:** The variables are stored in a database for further analysis.

Information analysis: Applying algorithms of data analysis and counting, the content of the variables are expressed in different reports that allow us to contrast the data and obtain relevant information.

2.1.1. Data universe

The methodology considers two main processes: **1) creation of directories and 2) obtaining the average weekly rate;** both processes are carried out to monitor the offer of traditional and non-traditional accommodations, using new technologies to exploit the information of the internet in a way that analyzes the behavior of the offer of the different types of accommodation.

The extraction of information is focused on the acquisition of data on a particular set of websites that provide the information required to determine the accommodation offer of each tourist destination. This is why the sources of information are reduced to online travel agencies (OTAs) and Metasearch engines, using the search criteria provided by these sites, obtaining data related to the accommodation offer of each tourist destination.

Currently, the hotel monitoring program of the Ministry of Tourism has directories obtained from work carried out by the governments of the states. The same program has information on 21 lodging variables in 70 destinations that it monitors weekly. This new methodology, based on Internet information, will support the updating of statistical information that is obtained today with the DataTur system for traditional accommodation.

The data obtained on non-traditional accommodation is an experimental set that will allow the sector to study the behavior of the shared economy. The methodology for non-traditional accommodation follows the traditional one and considers the 70 tourist destinations referenced by DataTur system.

The methodology is based on the use of a software tool to carry out the extraction and analysis of data and allows you to configure the monitoring periods and store the data chronologically. Also, the tool has a decoupled design that allows increasing the scope of information extraction, in such a way that tourist destinations and sources can be added with the necessary maintenance and support.

2.1.2. Variables

The implementation of the methodology requires a selection of variables of interest for the creation of accommodation directories (traditional and non-traditional) and the average weekly rate. The variables selected for the accommodation offered are described below:

1. Directory of traditional accommodation

• **Name:** Name of the accommodation by which they are offered in the OTAs.

• **Category:** It is based on the rating that each OTA gives to the accommodation, and may or may not coincide with the classification used by the Ministry.

• **The number of rooms:** The number of rooms that exist in the accommodation facilities according to the OTAs.

• **Location:** Address shown in the accommodation description.

• **Geo-referenced position:** Position in latitude and longitude available on the web page.

2. Directory of non-traditional accommodation

• **Name:** Descriptive name of non-traditional accommodation.

• **Type:** Type of accommodation, for example: a full house, shared room, apartment, among others.

• **The number of rooms:** Number of rooms offered by the establishment.

• **The number of guests:** Number of persons allowed to stay.

• **The number of beds:** Number of beds of the establishment.

• **Geo-referenced position:** Position in latitude and longitude.

3. An average weekly rate of Traditional accommodation: an Average weekly rate of tourist destinations by classification taking into account the accommodations available on the websites at the time of data extraction.

4. An average weekly rate of non-traditional accommodations: Average weekly rate for each tourist destination offered on the website, considering the non-traditional accommodations available at the time data was obtained.

2.1.3. Technique to obtain information

The technique used to obtain information consists in extracting the available information on websites that offer accommodations for tourist destinations in Mexico. From this extraction, a process is carried out to store the variables of interest and generate the diagnosis of the lodging offer. For this reason, the tool is made with information extraction techniques on the web, in such a way that the extraction will be done automatically.

Because the object of the study comes from the Internet and this is a source of information with a large amount of data, the extraction was defined to the OTAs and Metasearch engines that meet the required technological and information characteristics.

To perform the web search and extraction process automatically, a study was conducted to identify the parameters that must be entered in each site to obtain the list of accommodations, how many and which tourist destinations are included in each OTA, and which filters or criteria search should be applied to reduce the noise in the data.

In the case of obtaining a weekly average rate for traditional accommodation, a "standard room" for "two persons" and a one-week stay that includes the seven calendar days of "Monday to Sunday" is selected in the search for accommodation. This is with the purpose of being able to cross the information with the weekly reports by the DataTur system. In the case of non-traditional accommodations, the average weekly rate is extracted by selecting "two guests" and any type of accommodation, in regards to time, the guidelines are respected in traditional accommodations.

2.1.4 Periodicity of monitoring

According to the methodology, for the diagnosis of the average weekly rate and the accommodation offer, the periodic monitoring of the information of each tourist destination is required. To do this, with the help of the software tool you can define information extraction dates that will allow the creation of the directories and the statistical information of the average rates.

Directory of traditional and non-traditional accommodations: A single monthly extraction is scheduled for both directories, that is, they are executed sequentially, first the traditional sector and then the non-traditional sector.

An average weekly rate of traditional and non-traditional accommodation: Two extractions are made per month, indicating the target weeks as a reference to obtain the prices offered. The target weeks are established in the second and last complete weeks of the month to be extracted. The average of the weekly rate is made by a simple average with the total of the observations of each destination. These rates are not associated with the directory and do not consider the characteristics of the accommodation to weight such calculations, that is, each establishment contributes the same weight to the obtained rate.

2.1.5   Conflict identification

Based on the process established for directories, for the traditional sector, an additional process of unification of sources is carried out, in which the name and geographical position are established as parameters to associate accommodations and obtain a directory eliminating duplicates. In this case, when two lodgings present inconsistency in the data, the tool qualifies it as a conflict. The two types of inconsistency that the tool can detect are:

1. Similar names and long distances. E.g., Acamar Acapulco and Acamar Beach Acapulco, with a distance of 3 kilometers between them.

2. Different names and close distances. E.g., Acamar Acapulco and Costa Linda Acapulco, with a distance of 100 meters between them.

The tool presents the user with the list of conflicts that must be resolved with the assistance of an expert so that the directory can be stored, however, in the absence of this step, the tool has a mechanism that allows resolving conflicts, based only on the similarity of the name to associate or separate the accommodations, 24 hours before the next directory extraction, thus ensuring the construction of the directory.

2.1.6   Information analysis

This is the final stage of the methodology that consists of obtaining reports on the number of accommodations and the average weekly rate for each tourist destination. Therefore, the tool, which applies the methodology, processes the information extracted from the internet and stored in the database to generate different reports on the data.

These reports allow us to have information that models the online offer of traditional accommodations with which the behavior of the hotel industry on the Internet is modeled and can make comparisons with the information they generate with traditional sources. For non-traditional accommodations, it is experimental information with which this new business model can be analyzed in Mexico.

The reports that the tool can generate are:

- **Accommodation directories:** For both types of accommodation, traditional and non-traditional, you can consult the directories stored in the database, applying different search criteria, for example, the directory version.
- **Average weekly rate:** Report with the weekly average rate table, applying different filters, such as tourist destination, type of accommodation, among others.

An essential aspect of the implementation is to determine the periodicity of the monitoring of the information because a market in constant change is being studied, as is the rate and the accommodations offered, especially the one related to the shared economy or non-traditional accommodations. To do this, the technological tool has a parameter configuration module that allows you to set the extraction dates.

3. Results obtained

The object of the study was the 70 tourist destinations in Mexico that the Ministry monitors. The monthly directories of the hotel sector are generated from several sources of information (OTAs); the data that is extracted from the sources are processed by language processing algorithms that allow to eliminate redundant information and maintain a single directory. This information fusion is cooperative so that each of the sources contributes new accommodations to the monthly directory.

For the monthly directory of the extra-hotel sector and shared economy, data is extracted, and there is no change control; however it does have control of versions to indicate that it is the monthly one.

The average weekly rate extraction, both for the hotel and extra-hotel sector, is obtained from the Metasearch engines and Airbnb, respectively. The results show, mainly for rates, correspond to aggregate data statistics, in compliance with the legislation on privacy protection.

3.1 Results of data extraction for directories

The directories shown in this section were obtained from the extractions carried out in February, March, and April (2018), creating monthly directories for the hotel and extra-hotel sector.

3.1.1. Hotel or traditional accommodation directory

The monthly directory of traditional accommodation is obtained through the generation of a report with the software tool by selecting all or a single destination.

Description	
Summary table	
Total of hotels	Indicates the total number of hotels stored in the selected version, corresponding to the selected destinations.
Total of rooms	Indicates the total number of rooms obtained in the selected version according to the selected destinations.
New hotels	The number of new hotels obtained in the selected version, where it is verified that the date of creation is the same as the date of the selected version.
Change in hotels	The number of hotels that have had changes to the current date.
Hotels from the last version	The number of hotels that match the previous version to the current one, according to the selected destinations.
Data of the traditional accommodation records	
#	Indicates the number with which the record was stored in the BD of the software tool: is a numerical type of ID and is created consecutively.
Destination	Indicates the name of the tourist destination.
Name	Indicates the name of the accommodation proposed by the base source.
Category	Indicates the category of the accommodation, taking into account the maximum category established by the sources. It is a numerical type (floating).

Rooms	Indicates the total number of rooms in the accommodation, obtained from the website.
Geographic position	Indicates the latitude and longitude of the accommodation.
Address	Indicates the address of the accommodation provided by the source.
Date of creation	Indicates the version date in which that record was created in the BD of the software tool.

The extraction carried out in February, from the five OTAs, resulted in 6,915 hotels with a total of 577,027 rooms. In the month of March, 9,801 hotels were obtained with a total of 650,284 rooms; and for April 11,854 hotels were obtained with a total of 688,950 rooms.

With the results presented, it is observed that in each directory version the total of hotels, in the Hotels column was increased. However, there will not always be an increase in the number due to the fact that the total number of hotels corresponds to the accommodations that were offered on the date of the consultation. That is, it is possible that in other versions fewer hotels are obtained than those obtained the previous month. This depends on the stability and availability of the hotel offered by the OTA.

The total number of rooms (see the Rooms column) corresponds to the number of rooms with which the hotel is physically described in the OTA websites and does not refer to the number of rooms or beds available.

The New hotels column shows the number of hotels that appear for the first time compared to previous versions. As the data base grows, the number of new hotels decreases; this is due to the mechanisms of control of changes and control of versions, existing in the database, which allow to store a single record and maintain historical updates.

Based on this data we can observe behavior related to the inclusion of hotels in the OTA websites, which may be new or entering only in certain season to be offered. Therefore this business model is constantly changing.

Date of extraction	Hotels	Rooms	New hotels	Hotels from the last version
2018-02	6,915	577,027	6,915	0
2018-03	9,801	650,284	3,835	5,966
2018-04-I	11,206	650,763	2,622	8,198
2018-04-II	11,854	688,950	981	10,103

Destination	02/2018		03/2018		I/04/2018		II/04/2018	
	# hot.	# rooms	# hot.	# rooms	# hot.	# rooms	# hot.	# rooms
Acapulco	134	14,619	154	15,267	173	18,717	148	17,372
Cancun	212	32,569	193	22,520	223	26,897	259	34,915
Guadalajara	181	15,434	183	15,905	198	15,735	185	16,161
Mazatlan	101	9,647	98	9,432	100	9,368	94	8,237
Monterrey	106	13,114	107	12,684	117	13,095	116	13,109

In contrast, there are tourist areas with a lot of variation between versions on the number of rooms ± 200. This is because the OTAs offer accommodations that are outside the geographical coverage of the destination. That is to say, they incorporate destinations such as Los Cabos, which may include hotels in Cabo San Lucas, Los Cabos Corridor Area and San Jose del Cabo. For this type of problems, it is proposed to integrate a filtering of accommodation by georeferenced area. Likewise, it is a variable market that allows to observe some hotels bid only in certain periods, depending on their marketing strategies.

Name	Cat.	Rooms	Geo position	Address	Date of creation	Status	Last Change
Hotel 1	5	324	16.8561,-99.8662	COSTERA MIGUEL ALEMAN	14/02/2018	MODIFIED	23/03/2018
Hotel 2	5	23	16.8113,-99.8479	CARR. ESCENICA KM 14 FRACC. PICHILINGUE	14/02/2018	WITHOUT MODIFICATIONS	14/02/2018
Hotel 3	5	61	16.7924,-99.8269	PASEO DE LA QUINTA # 6	14/02/2018	WITHOUT MODIFICATIONS	14/02/2018
Hotel 4	4	257	16.8576,-99.8683	AV. COSTEA MIGUEL ALEMAN	14/02/2018	WITHOUT MODIFICATIONS	14/02/2018

3.1.2. Non-traditional and extra-hotel accommodation directory

Each monthly directory of the extra-hotel sector has the 70 destinations taken from Airbnb. The non-traditional sector does not have regulations or background of its behavior.

Description	
Summary table	
Total accommodations	Indicates the total number of accommodation stored in the selected version, corresponding to the selected destinations.
Total rooms	Indicates the total number of rooms obtained in the selected version according to the selected destinations.
Data of non-traditional accommodation data	
#	Indicates the number with which the record was stored in the BD of the software tool: is a numerical type of ID and is created consecutively.
Destinations	Indicates the name of the destination
Name	Indicates the name of the destinations proposed by Aribnb.
Type	Indicates the type of accommodation: shared room, apartment or house
Rooms	Indicates the total number of rooms of the accommodation
Beds	Indicates the number of beds of the accommodation
Geographic position	Indicates the latitude and longitude of the accommodation

Date of extraction	Total accommodations	Total rooms
2018-02	13,874	23,457
2018-03	14,852	24,542

2018-04-I	14,830	24,382
2018-04-II	12,185	20,097

3.2. Rates

Another important data obtained with the software tool is the average rates that the accommodations offer on the websites. The average weekly rate of each accommodation is considered to calculate the minimum, maximum, average and standard deviation values that are provided for each resort. The following table describes the fields contained in the rate report.

	Description
Destination	Indicates the name of the destination
Number of data	The total data of the week
Minimum rate	It is the minimum rate for lodging of the data set obtained for the tourist destination in the selected version.
Maximum rate	It is the maximum rate for lodging of the data set obtained for the tourist destination in the selected version.
Average	It is the average rate calculated from the accommodations set obtained for the tourist destination in the version selected for the report.
Standard Deviation	It is the standard deviation calculated with the rates of the accommodations set.

To calculate the average, the maximum and minimum value of the data set obtained are eliminated; and the null or empty values are not considered in these calculations.

3.2.1. Hotel rates

The tool allows the generation of rates reports in summary mode and as a percentage variation between versions. In the rate summary report, two types are generated: by version and by month. In the month type, the conglomerate of the two objective weeks extracted during the month is described.

The unification process includes the following rules:

1. If it is the same name and it is the same cost then it is unified.
2. If it is the same name and has different cost then they are taken as different observations.

Therefore, the number of observations is less than or equal to the sum of the observations of the two versions.

Destination	Number of data	Minimum rate (pesos)	Maximum rate (pesos)	Average	Standar deviation
Acapulco	368	$290.00	$10,500.00	$1,539.35	$1,395.83
Aguascalientes	101	$493.00	$5,715.00	$1,662.73	$938.81

Akumal	88	$639.00	$10,931.00	$3,190.83	$2,281.31
Huatulco	175	$336.00	$5,377.00	$1,254.47	$929.52
Cabo San Lucas	259	$485.00	$124,855.00	$8,897.24	$14,930.70

Average rates by type of center figures in national currency (mxn)							
Type center	April		May		June		July
Borders	$	1,910	$	1,503	$	1,107	$ 1,114
Big cities	$	2,814	$	1,464	$	1,435	$ 1,463
Fully Planned	$	4,830	$	4,655	$	3,458	$ 3,440
Cities of the interior	$	1,375	$	1,468	$	1,000	$ 1,198
Others	$	2,983	$	2,655	$	2,210	$ 2,291
Traditional beach	$	2,651	$	2,667	$	1,531	$ 1,836

Source : Ministry of Tourism 2018
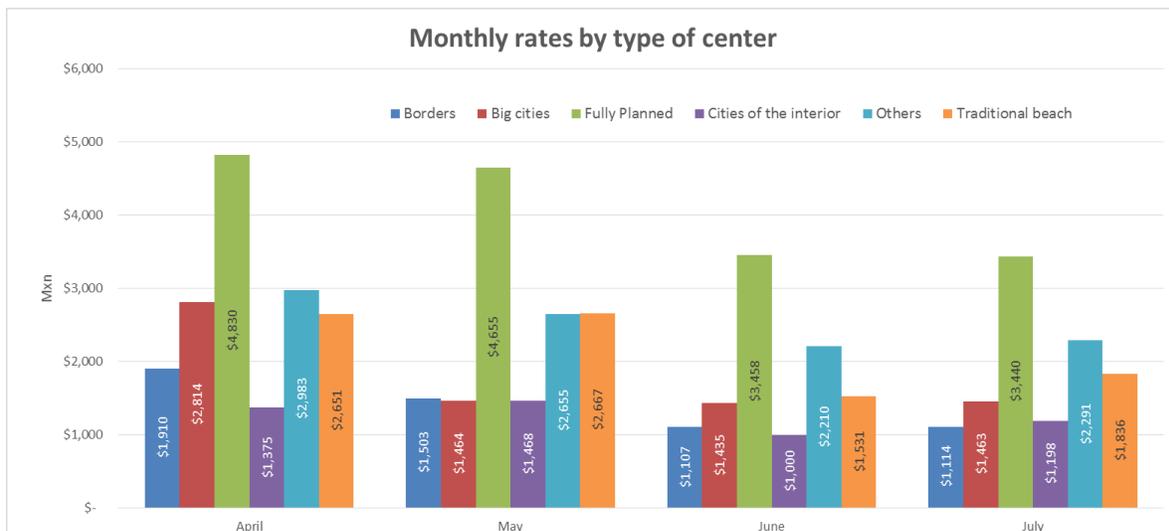
It is important to mention that all reported rates do not include taxes.

Some of these centers remain within the first 10 places in both versions and it is possible to observe that the target week has an impact on the prices of the rooms. It is important to mention that tourist destinations such as Los Cabos show a very separate range in the number of observations, this can be the cause of:

Incorporation of destinations by proximity. For example, Los Cabos, on occasions, as a result of the search, the OTA includes hotels in Cabo San Lucas, the Cabo Corridor Area and San Jose del Cabo.

1) Incorporation of houses and villas in the category of hotels, so the increase in the offer is affected.

April/2018			April/2018		
Destination	# Obs.	Average (pesos)	Destination	# Obs.	Average (pesos)
Cabo San Lucas	157	$11,261.7	Playacar	89	$15,132.4
Los Cabos Corridor	246	$5,238.5	Los Cabos Corridor	245	$9,385.79
Los Cabos	193	$4,772.38	Mexico City	661	$5,886.27

San Jose del Cabo	91	$4,034.07	San Miguel de Allende	137	$5,221.99
Querétaro	185	$3,962.26	Cabo San Lucas	105	$5,137.26
Nuevo Vallarta	47	$3,668.3	San José del Cabo	40	$5,116.8
La Paz	230	$3,617.07	Los Cabos	83	$5,041.11
Akumal	38	$3,486.84	Tijuana	98	$4,933.65
Cozumel	333	$3,363.58	Valle de Bravo	34	$4,751.5
Playa del Carmen	473	$3,142.29	La Paz	158	$4,553.63

4.   Conclusions

The challenge of tourism in Mexico is to increase competitiveness through public policies and programs such as "Viajemos por México" (Let's travel through Mexico), "Mejora tu Hotel" (Improve your Hotel) and "Conéctate al Turismo" (Connect to tourism). These actions must arise from a diagnosis about the current state of tourism that will support the design and implementation of these policies. Therefore, the importance of strengthening the information that supports the development of policies; improving the use, integration and availability of tourism statistics to guide the creation of such policies, guide their implementation, support monitoring and evaluation.

Given the need to have information that supports decision making, one of the objectives of this project was the creation of a practical methodology for the generation of statistics. The methodology is based on the methodological document of the Mexican National Institute of Statistics and Geography (INEGI) consumer price index, specifically on the considerations that the institute takes into account in order to consult about the price per room.

Currently, with the technological tool, the accommodation directories and the average rates of the 70 tourist destinations have been obtained for the months of February to September 2018. It is important to mention that, based on the information generated with the tool, it will be possible to monitor and analyze the behavior of the hotel and extra-hotel sector regarding the number of accommodations and the weekly average rates offered on the Internet. Also, to classify the tourist centers based on their average rates and in future compare them with the international market. The strategy used in the proposed solution was the integration of several information sources to merge them in a complementary manner.

Likewise, with the data obtained from the lodgings of each tourist center, it will be possible to generate a more representative sample; have the georeferenced information to represent the accommodations in the website and mobile application "ATLAS Turístico" (Mexico Tourist Atlas – the systematic recording of all public property, natural and cultural sources that may become national attractions, places of interest and generally all those areas and territorial areas) while minimizing the times of obtaining georeferenced data.

The statistics of this methodology obtained from the web sites present temporary dynamism, due to the fact that hotels and accommodation owners do not publish the offer throughout the year on these websites and can be seen in the results obtained. Also, we observe the variation in rates depending on the season and this was one of the most important challenges to solve in this investigation, the constant change of websites not only in their content but in the structure of the web pages; so the data extraction techniques of the Internet must be robust to these changes.

With the results obtained in this project, it was possible to confirm that the technological enablers significantly help the Ministry of Tourism, to be at the forefront on the analysis of the information

generated by tourists and suppliers. The exploitation of this information, in a timely manner, will allow the Ministry to make public policy decisions related to the accommodation offer of both hotel and extra-hotel sectors.

Finally, this paper identified the future work of the housing offer monitoring methodology, which is listed below:

- The design and implementation of an accommodation data cleaning algorithm based on the georeferenced coverage of each resort.
- The extraction of data on the services offered by each accommodation to make a classification by type of services.
- The analysis of other sources of information on shared economy and incorporate the fusion of several sources.
- The design and implementation of the physical and logical architecture to extend the extraction of data contained in the websites of interest, without limiting the consultation to the established in the methodology (single room for two people), as well as integrating more tourist centers in Mexico and other countries.
- The comparative analysis between the sources of the Ministry of Tourism, INEGI and the technological tool to identify the percentage of the market that is commercialize on the Internet by the total registered.

5.   References

[1]   C. M. M. E. P. a. O. L. I. Amaya Molinar, «Evaluación de destinos turísticos mediante la tecnología de la ciencia de datos,» de Estudios y perspectivas en turismo, 2017.

[2]   CONACYT,   «http://conacytprensa.mx/index.php/tecnologia/tic/9592-desarrollan-sistema-de- informacion-turistica,» [En línea].

[3]   Z. S. Z. G. J. J. H. a. U. M. Xiang, « What can big data and text analytics tell us about hotel guest experience and satisfaction?,» de International Journal of Hospitality Management, 2015.

[4]   Z. D. Q. M. Y. a. F. W. Xiang, «A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism,» de Tourism Management, 2017.

[5]   R. a. B. R. Oliveira, «Extracting web data from TripAdvisor as a support for tourism indicators development in Minas Gerais,» de 14th Global Forum on Tourism Statistics, Venice, Italy, 2016.