GOBIERNO DE ESPAÑA
MINISTERIO DE TRANSPORTES Y MOVILIDAD SOSTENIBLE

QUALITY CONTROLS PERFORMED ON MOBILITY ANALYSIS WITH BIGDATA

Part 1. Basic studies

Version 1.0. October 2024

Subdirectorate General for Planning, Trans-European Network and Logistics

General Directorate of Mobility Strategies


GOBIERNO
DE ESPAÑA
MINISTERIO
DE TRANSPORTES
Y MOVILIDAD SOSTENIBLE

Content

1. Purpose of the document. Previous considerations

This document makes a compilation of all the validations performed in the mobility study with Big Data from the Ministry of Transport and Sustainable Mobility to assess the reliability of the data generated in the basic products (basic travel matrix and matrix of overnight stays). This also constitutes valuable information for its potential users, since it allows them to know the degree of accuracy in its different aspects for use in various applications.

Real mobility cannot be characterized with exact metrics, it can be said that it is an unknown measure, so we cannot directly compare the study data with an absolute reference standard that allows us to assess its reliability. However, it is possible to carry out controls that compare the study data with each other and with other certain reference data in terms of variations and trends, which allow us to detect the presence of errors and assess the reliability of the data.

Another aspect to keep in mind about these validations is that, due to the very characteristics as a BigData project, it is not feasible to evaluate each data individually, but rather aggregate or general analyses are performed to obtain an overall assessment of the data.

Since the beginning of the project, the data have undergone numerous controls and validations of various kinds, both periodic validations on the data received, as well as punctual or visual validations, or those that evaluate the consistency and logic of the data. This document will present those performed to evaluate the matrices of the basic study. It is important to note that these validations join a long list, among which are those developed to obtain an initial validation of the methodology, or those performed periodically to ensure the quality of the telephony data received as input for the whole process, both performed by Nommon.

2. Validations to the basic travel matrix

2.1- Continuous automatic control. Anomaly detection

This validation is a continuous automatic control applied to the basic trip matrices upon receipt by the Ministry of Transport and Sustainable Mobility, in order to detect anomalies in the processed data by detecting outliers (atypical, extreme or extraordinary values) with respect to the time series of data.

During this automatic process only anomalies are detected, which must be evaluated a posteriori, since anomalies in the data can come from errors in the initial data (lack of telephone records in some areas, antenna geolocation errors, etc.), as well as from real mobility anomalies, such as those caused by holidays, mass events, changes in the weather or other mobility disrupting events.

Currently, based on the analysis of the results of these automatic controls, the mobility anomalies are justified and the type days and singular days are determined, which will be used to choose the days on which the complete matrices will be generated.

In phases prior to this study, these checks also allowed the detection of errors from the mobile operator, such as problems with antennas, which occurred with some frequency before the methodology was debugged in 2022. Thanks to Nommon's implementation of pre-checks on incoming record volumes, errors are detected as early as possible so that data can be recovered if necessary, so this is an issue that does not happen today. In the current study, 2022-2024, the only incidence of raw data from Orange, where data could not be subsequently recovered, occurred in the months of October and November 2023, which was detected by Nommon when the records for those days were received.

This validation, implemented using ETL technology, is done on each study day, comparing the trips of that day with those of the same day of the week in the previous four weeks, and with those of a reference date of a typical week.

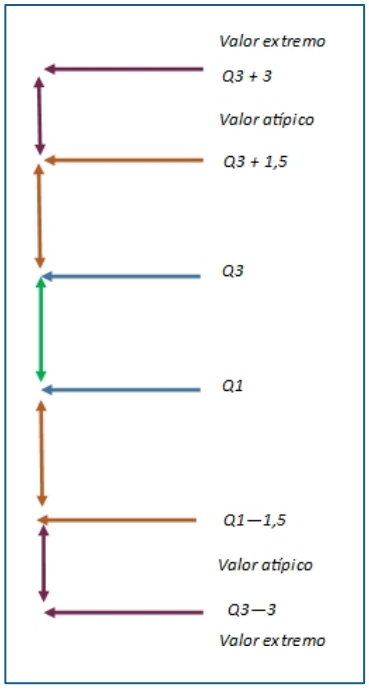
It consists mainly of a comparison made by means of zonal aggregations at various levels, to which different criteria are applied for the evaluation of the acceptance of the study day, the aggregation zones being the following:

- Inter-municipal aggregation
- Intra-municipal aggregation
- Total-municipal aggregation
- Inter-provincial aggregation
- Intra-provincial aggregation
- Total-provincial aggregation
- Provincial aggregation-OD

The comparison of the study day with reference days measures discrepancies in travel between dates and evaluates these discrepancies using preset thresholds, some of which are a function of statistics on the data:

Threshold	Value	Threshold adjusted with the quartiles Q1 and Q3	
Atypical threshold	1,5	High outlier threshold	*Q3+Umbral atypical x (Q3-Q1).*
		Low outlier threshold	*Q1-Atypical threshold x (Q3-Q1).*
Extreme threshold	3	Extreme high threshold	*Q3+Umbral end x (Q3-Q1).*
		Threshold extreme under	*Q1-Extreme threshold x (Q3-Q1)*
Maximum variation average	5		
Maximum variation average alarm	15		
Minimum factor Reference	25		
Maximum factor Reference	125		
High alarm threshold	110		
Threshold under alarm	90		

The following diagram refers to the different thresholds considered:



During this analysis it is determined, for each zone, if the variation in the number of trips with respect to the reference dates is within the established thresholds, or if not, it classifies the aggregations in the following anomalies:

Pattern comparison	Aggregation	Comparison	Anomaly
Previous 4 weeks	Municipal	Travel < low outlier threshold	Atypical low
		Travel > high outlier threshold	Atypical high
		Travel < extreme low threshold	Low end
		Travel > extreme high threshold	High end
	Provincial	Travel < low outlier threshold AND Average trip variation >5 or <15	Atypical low
		Travel > high outlier threshold AND Average trip variation >5 or <15	Atypical high
		Trips < low extreme threshold AND Average trip variance < Variance. minimum average alarm	Low end
		Trips > high end threshold AND Average trip variation > Variance maximum average alarm	High end
		Travel > extreme high threshold AND Average trip variation > Average max. variation alarm	High alarm
		Travel < extreme low threshold AND Variation of average trips < Variation min average alarm	Low alarm
Date of Reference	Municipal and Provincial	Travel/TravelReference > Maximum factor Reference	Exceptional high

		Trips/TravelReference < Maximum factor Reference	Exceptional bass
		Travel/ReferralTravel > High Threshold Reference	High alarm
		Travel/TravelReference < Low Threshold Reference	Low alarm

Once each aggregation zone has been classified, the study day is evaluated by meeting the criteria defined for each type of aggregation:

Aggregation	Acceptance criteria
Municipal Total	5% of areas with outliers or extremes are allowed.
Municipal Intra	There can be no outliers or extremes, unless justified.
Municipal Inter	5% of areas with outliers or extremes are allowed.
Provincial Total	Zones with alarms are not allowed. 5% of zones with outliers or extremes are allowed.
Provincial Intra	Zones with alarms are not allowed. There can be no outliers or extreme values, unless justified.
Provincial Inter	Zones with alarms are not allowed. 5% of zones with outliers or extremes are allowed.
Provincial OD	5% of areas with outliers or extremes are allowed.

After this evaluation, the typical days and the anomalous or singular days are extracted, and it is verified that they correspond to real anomalies, i.e., that there has been some event that justifies the change in the mobility pattern, which occurs in 99% of the days detected with anomalies.

2.2- Evaluation of anomalies using GIS and BI

These controls are carried out on a more or less continuous basis depending on the time of the project's evolution in order to detect and analyze mobility anomalies.

Visual review with GIS technology

This review makes it possible to evaluate the anomalies detected in the previous quality control with a geospatial vision. The results of the quality controls are stored on a server (SQL server) and from a visualization service it is possible to analyze the anomalies in a global way.

When there are spatially clustered anomalies it could be due to errors in the data source, due to lack of synchronization with the antenna map (these errors were relatively frequent until 2021) and also due to local festivities or other mobility disrupting events. For this reason, anomaly detection can be automatic but its evaluation cannot be 100% automated.

Extraction of typical days and singular days

Currently, based on the analysis of the results of the automatic controls, the mobility anomalies are justified and the typical days and singular days are classified, which will be used to choose the days on which the complete matrices are generated.

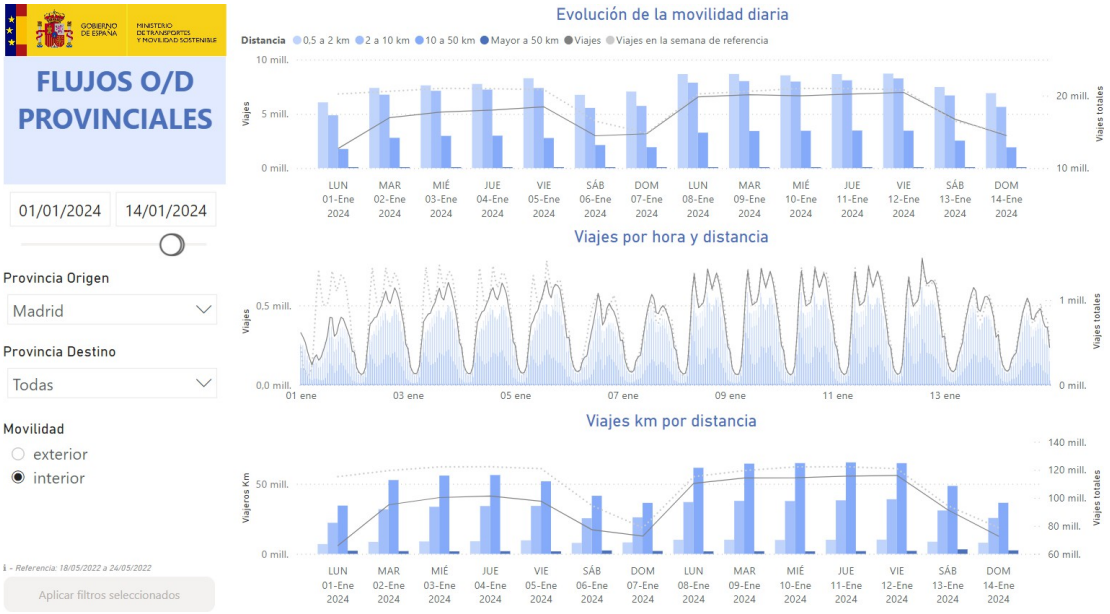
	L	M	X	J	V	S	D
						1 (D.S.)	2 (D.S)
Enero	3	4	5 (D.S.)	6 (D.S.)	7 (D.S.)	8 (D.S.)	9 (D.S.)
	10 (D.S.)	11	12	13	14 (D.S.)	15 (D.S)	16
	17	18	19	20	21	22	23
	24	25	26	27	28	29	30
Febrero	31	1	2	3	4 (D.S.)	5 (D.S.)	6 (D.S.)
	7	8	9	10	11	12	13
	14	15 (D.S.)	16 (D.S.)	17	18	19	20
	21	22	23	24 (D.S.)	25 (D.S.)	26 (D.S.)	27 (D.S.)
Marzo	28 (D.S.)	1	2	3	4 (D.S.)	5	6
	7 (D.S.)	8	9	10 (D.S.)	11 (D.S.)	12 (D.S.)	13 (D.S.)
	14 (D.S.)	15	16	17	18 (D.S.)	19 (D.S.)	20
	21	22	23	24	25	26	27 (D.S.)
	28	29	30	31	1	2	3
Abril	4	5	6	7	8 (D.S.)	9	10 (D.S.)
	11	12	13 (D.S.)	14 (D.S.)	15 (D.S.)	16 (D.S.)	17 (D.S)
	18 (D.S.)	19	20	21	22	23	24
	25	26	27	28	29 (D.S.)	30 (D.S.)	1 (D.S.)
Mayo	2 (D.S.)	3	4	5	6	7	8
	9	10	11	12	13 (D.S.)	14 (D.S.)	15 (D.S.)
	16 (D.S.)	17	18	19	20	21	22
	23	24	25	26	27 (D.S)	28 (D.S.)	29 (D.S.)
	30 (D.S.)	31	1	2	3 (D.S)	4 (D.S.)	5 (D.S.)
Junio	6 (D.S.)	7	8	9	10	11	12
	13	14	15	16 (D.S.)	17	18	19
	20	21	22	23 (D.S.)	24 (D.S.)	25 (D.S.)	26 (D.S.)
	27 (D.S.)	28	29	30	1 (D.S.)	2 (D.S)	3 (D.S.)

Each anomaly is justified in an anomaly file.

Graphical review in PowerBI

Sometimes, in the task of evaluating some anomaly previously detected by automatic processes and subsequently verified as explained above by GIS, a detailed review is required and is performed in PowerBI. For example, the time profile can be checked, which allows to deduce whether that day has been a holiday or not.

(analyzing peak commuting hours) and if there has been any peak time that indicates the presence of specific mobility disruptive events such as a soccer match at a certain time or any other mass event.



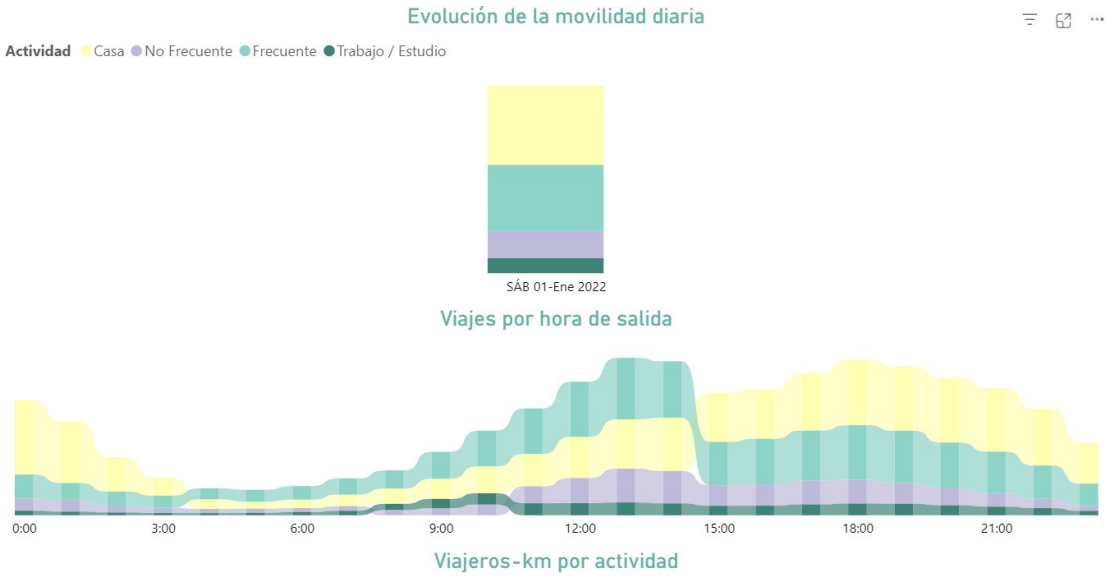
The PowerBi tool on the Mobility with Big Data website[1] allows you to make a visual comparison of daily mobility data with data from a reference period, allowing you to detect matches or discrepancies with expected patterns or to see trends, for example. It allows to evaluate the annual, daily and hourly evolution.

With this tool, the rest of the attributes of the OD matrix, such as patterns by distance, age, sex, income or residence, are reviewed in a visual, non-systematic way, being able to detect abrupt changes in the patterns by income, for example, which could be evidence of an erroneous crossing with the INE data, or unexpected patterns in any of the variables that would allow us to detect possible errors in some applied process or in the data.

For example, the following image illustrates the trend of number of trips per hour and age according to the expected behavior pattern:

[1] Basic studies | Ministry of Transportation and Sustainable Mobility

Or in the following one, which, through symmetry perceived in the morning/afternoon schedules for the development of home/work activities, shows the logical consistency existing in the mobility segmented by "departure time" and "activity":



Or by segmentation by "time of departure" and "income level":

This check is available to any user through this website, and the mobility pattern can be consulted for any combination of date, zoning and segmentation of traveler characteristics.

2.3- Internal consistency. Evaluation of symmetry

The symmetry of the OD matrices represents the balance between outbound and return trips to reciprocal origins and destinations on the same day. If all travelers made the same return trip as the outward trip, the OD trip matrix would be totally symmetric, therefore, the asymmetry of this matrix indicates the difference between that balance. We do not expect perfectly symmetric matrices, but rather matrices with a high coefficient of symmetry, since it is to be expected that there will be asymmetries due to variations in the place of overnight stay or to the existence of travelers who do not make exactly the same return trip as the outbound trip.

This symmetry evaluation has been applied punctually to a selection of matrices, with different aggregations, to check the behavior in each case, verifying a generalized symmetry.

The results obtained with respect to the symmetry of the matrices are highly satisfactory. The results obtained are shown in *the Appendix Matrix symmetry*.

2.3- Validation against reference data

The rate of change in the number of trips abroad in the years 2022 and 2023 obtained with BIGDATA has been compared with those provided by FAMILITUR. As absolute values of trips cannot be faithfully compared between the two sources, since they do not measure identical concepts, the analysis focused on comparing the rate of change of trips between two consecutive years.

To obtain this figure according to FAMILITUR, the figure for the concept of "trips", which includes trips with at least one overnight stay, has been added to that of "excursions", which are day trips and do not require an overnight stay, excluding those made frequently and those made by travelers under 15 years of age.

To obtain this figure according to BIGDATA, the interactive Monthly Mobility panel of the BIGDATA Mobility WEB was consulted.

FAMILITUR

Figures in Millions	**2022**	**2023**
travel	16	19,3
excursions	2,5	3
total	18,5	22,3

FAMILITUR-BIGDATA COMPARISON

(Amounts in Millions)	2022	2023	INTER-ANNUAL VARIATION RATE
FAMILITUR	18,5	22,5	**122%**
BIGDATA	54,3	64,2	**118%**

As indicated at the beginning of this section, since we are not measuring identical concepts and the measurement methodologies are totally different, we can observe a high disparity between the results obtained with both sources. However, there is a high correlation between the annual variation rate of both sources, so the results are considered satisfactory.

3. Validations to the matrix of overnight stays

3.1- Logical consistency

Based on the matrix of overnight stays, a matrix showing the reciprocity between the areas of residence and overnight stays for a typical day, October 18, 2022, has been generated, which is shown in the Annex *Matrix of Relationship between Zone of Residence vs. Zone of Overnight Stay,* in which a series of characteristics consistent with the expected logic can be observed*:*

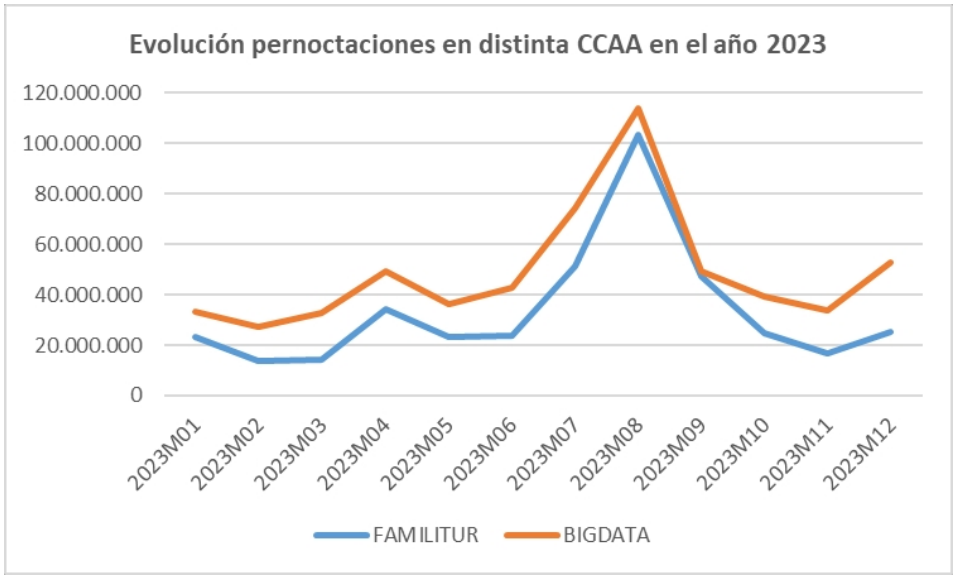
- Much higher values are observed on the diagonal, indicating that the number of people who spend the night at the same place of residence is much higher than those who spend the night at a different place.

- Without being a symmetrical matrix, since there are places that are more attractive for overnight stays and there is a certain reciprocity in the order of values, which indicates that the Autonomous Communities with the highest number of residents are those that host the highest number of overnight stays.

- The sum of the number of overnight stays for each of the columns of this matrix represents the sum of residents per Autonomous Community according to the study, which, when compared with the INE figures for residents per Autonomous Community, shows a very high correlation of values.

3.2- Validation against reference data

The evolution of the number of overnight stays in different Autonomous Communities over the year 2023 obtained with BIGDATA and FAMILITUR has been compared, showing a discrepancy, not very high, which can be attributed to several factors such as, for example, ambiguity in the area of habitual residence. However, the very high parallelism of the time evolutions of both series is verified, so the results are considered satisfactory:


Evolución pernoctaciones en distinta CCAA en el año 2023
120.000.000
100.000.000
80.000.000
60.000.000
40.000.000
20.000.000
0
2023M01
2023M02
2023M03
2023M04
2023M05
2023M06
2023M07
2023M08
2023M09
2023M10
2023M11
2023M12
FAMILITUR
BIGDATA

ANNEXES

Annex Matrix symmetry

Symmetry of travel between ACs on a typical day

Below is an OD matrix of trips between different ACs on a typical day on October 10, 2022.

	Comunitat Valenciana	Community of Madrid	Principality of Asturias	Basque Country/Euskadi	Extremadura	Cantabria	Castile-La Mancha	Andalucia	Region of Murcia	La Rioja	Navarre	Galicia	Illes Balears	Canary Islands	Catalonia/Catalonia unya	Melilla	Aragon	Castile and Leon	Ceuta
Valencian Community	14720570	13053	369	1250	446	308	22524	6154	91299	286	846	664	2268	654	25954	4	9207	1744	60
Community of Madrid	11249	20590404	1977	3604	4870	1528	301900	14749	2549	925	1448	4370	2023	3683	11366	97	4210	43720	23
Principality of Asturias	254	2342	2833306	1243	114	7428	305	280	89	75	145	14278	85	129	470	7	93	6511	
Basque Country/Euskadi	1172	4209	1137	6480024	224	33104	486	1046	113	21369	25419	679	267	797	1944		1880	25090	6
Extremadura	390	5587	111	370	2249406	47	5545	11727	138	30	42	248	57	109	192		151	7743	
Cantabria	317	2012	7496	33499	63	1709250	198	301	8	380	354	311	72	170	444		334	8710	
Castilla-La Mancha	22493	307174	336	482	5283	238	4335565	9811	12609	210	465	641	156	292	2080	5	1759	8145	
Andalucía	6284	15680	228	1272	11778	276	9890	23265385	22228	137	294	985	1111	2120	4730	525	744	1973	2258
Region of Murcia	91423	3163	97	154	192	34	12407	22255	4542405	103	202	263	162	224	1395	2	412	452	17
La Rioja	289	941	58	21915	25	464	261	121	80	809230	39051	133	4	30	660		2332	4864	
Navarra	714	1515	102	25348	13	320	406	195	50	39237	1751316	77	22	68	1443		13312	2011	
Galicia	579	5265	14327	838	297	429	732	1147	201	126	141	8162161	147	1115	1228		305	10496	
Illes Balears	2364	2508	48	446	110	81	220	1439	220		11	168	3743021	346	4057	9	104	233	
Canary Islands	633	3622	249	569	137	115	308	1565	175	29	70	1157	157	7242521	1535	3	11	309	
Catalonia/Catalonia unya	25176	11602	453	2117	151	306	2550	4622	1221	779	1468	1050	3612	1721	24052070	13	30826	1448	16
Melilla	6	57				2		408	48				13	10	14	277200	2	13	
Aragon	9188	4585	126	1928	98	324	1644	740	256	2122	13823	252	50	83	30638	4	3417688	4248	
Castilla y León	1632	45869	6640	25665	7422	8925	8456	2003	474	4735	2288	10717	252	352	1506		4212	6012823	3
Ceuta	3	18			2		5	2193	7					5	13	17		3	261243

From this matrix *M*, the difference matrix with respect to its transpose, matrix *D,* has been calculated*,* which will have higher values the more asymmetric the matrix *M* is*,* obtaining the asymmetry coefficient by means of the ratio between the Frobenius norm of both matrices*:*

$$Coeficiente\ de\ asimetría = \frac{||D||_F}{||M||_F} = \frac{\sqrt{\sum_{i,j}\left(M_{i,j} - M_{j,i}\right)^2}}{\sqrt{\sum_{i,j} {M_{i,j}}^2}}$$

This coefficient gives a value in the range [0,1], with 0 being a perfectly symmetric matrix and 1 being a fully asymmetric matrix.

Calculating this asymmetry factor we obtain a value of **0.0002**, which indicates an almost perfectly symmetrical matrix. As we said, we do not expect perfectly symmetrical matrices, but rather matrices with a high symmetry coefficient, since certain asymmetries are expected due to variations in the place of overnight stay, or also due to the existence of travelers who do not exactly replicate the outbound trip in the return trip.

Symmetry of trips between municipalities in Madrid on a typical day

Below is an extract of an OD matrix of the trips between the different municipalities of Madrid on October 10, 2022, in which, despite not being totally symmetrical, there is a high correlation between both sides of the diagonal:

	28002	28004	28005	28006	28007	28008	28009	28010	28012_AM	28013	28014
28002	2795		2238	316	39		210	7	15	14	114
28004		5071	23	13	582		3			6	6
28005	2292	24	435042	1524	705		627	51	1508	185	1318
28006	330	10	1522	183838	1117	4	3137	109	28	163	397
28007	44	671	605	1155	272388	180	122	104	23	341	385
28008		7		4	163	684	3	4		16	22
28009	193	7	648	3009	133		13238	6		7	76
28010			54	95	109	4	10	6443		7	24
28012_AM	9		1495	32	29		5		442	7	380
28013	13	6	131	146	346	16	5	29	7	110209	148
28014	99	9	1377	381	498	2	92	6	353	151	102504

Calculating this asymmetry factor we obtain a value of **0.0006**, which also indicates an almost perfectly symmetrical matrix.

Annex matrix of relationship between zone of residence vs. zone of overnight stay

This matrix shows the relationship between the number of people residing and staying overnight in each Autonomous Community, from which the conclusions set out in section 3.1 of Logical Consistency of the matrix of overnight stays can be drawn:

	Valencian Community	Community of Madrid	Principality of Asturias	Basque Country/Euskadi	Extremadura	Cantabria	Castilla-La Stain	Andalucía	Region of Murcia	La Rioja	Navarra	Galicia	Illes Balears	Canary Islands	Catalonia/Catalonia	Melilla	Aragon
Community Valenciana	4956210	18732	1941	4372	1655	1480	10746	13216	7801	810	1565	3722	4128	2897	14659	99	6378
Community of Madrid	27425	6537226	5129	5800	8884	4382	42035	32154	5965	1308	2408	11599	5096	9799	15927	309	5288
Principality of Asturias	2983	4813	980754	1438	571	1398	756	2891	393	215	285	4192	562	1806	1863	23	617
Basque Country/Euskadi	7272	5408	1572	2147208	1581	5894	923	5239	564	2935	4132	3574	1533	4079	4720	24	1882
Extremadura	1359	8317	599	2013	1023796	415	2384	7692	397	215	234	792	511	680	3020	41	815
Cantabria	1757	2580	1130	4267	306	563939	463	1601	145	316	397	1133	380	1017	1043	17	509
Castilla-La Mancha	13194	33006	1073	1216	2664	687	1952579	10480	3564	415	565	2059	1262	1514	5725	61	2343
Andalucía	12990	28542	3289	4664	8372	2161	10415	8331401	8564	1129	1615	6646	6413	7492	20667	1174	4188
Region of Murcia	10607	5256	415	607	493	274	3265	7917	1480017	211	423	1367	879	803	3509	52	761
La Rioja	1763	1340	151	2729	210	322	353	1335	196	303383	2198	340	69	258	1069	4	1113
Navarra	2554	2241	288	4153	295	589	539	2478	313	1519	627247	531	231	560	2215	2	2244
Galicia	2871	8873	3672	2814	847	971	1340	4414	640	429	549	2648454	1392	5139	5260	84	1085
Illes Balears	4576	4712	611	1019	703	396	1036	7068	720	83	264	1424	1139048	1448	7686	16	667
Canary Islands	2476	7131	1310	1452	707	579	928	6776	676	209	375	3184	1339	2135205	3990	191	765
Catalonia/Catalonia	14722	14090	1653	4217	2396	1347	4167	19086	2498	1123	2098	6074	6789	5974	7658439	151	10029
Melilla	354	416	34	93	77	21	64	2583	175	18	19	61	83	208	335	81440	101
Aragon	6071	4195	624	1737	602	550	1861	4001	655	820	1944	1149	578	1158	9684	54	1277625
Castilla y León	9146	21275	4667	9353	2916	3576	2972	7986	1126	1355	1404	6573	1339	3032	5878	64	2993
Ceuta	175	366	18	105	60	23	73	3012	113	10	56	56	24	88	181	75	33
COLUMNAR SUM	5078504	6708518	1008930	2199255	1057137	589004	2036897	8471331	1514523	316504	647779	2702931	1171659	2183156	7765871	83882	1319435
Residents according to INE	5186582	6851312	1005533	2212450	1054682	587527	2076182	8564311	1545698	321679	670606	2697287	1206684	2205349	7868200	85272	1337461