Regionalising tourism statistics

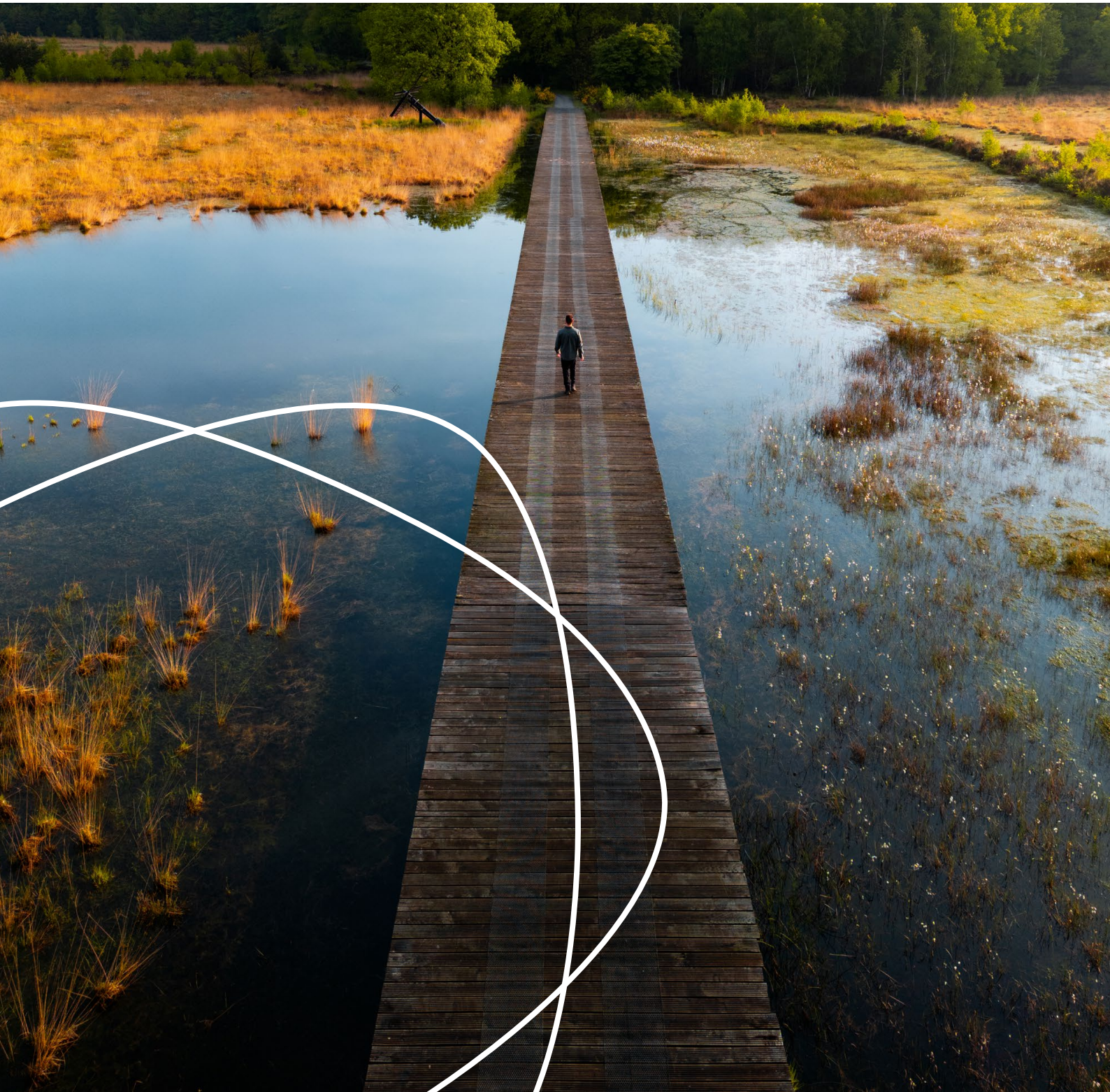
Report Data & Development Lab

CELTH Centre of Expertise leisure, tourism & hospitality

NBTC Nederlands Bureau voor Toerisme & Congressen

HZ UNIVERSITY OF APPLIED SCIENCES

Centraal Bureau voor de Statistiek



Colophon

This report was compiled within the Data & Development Lab (DDL), a partnership of Statistics Netherlands (CBS), Netherlands Board of Tourism & Conventions (NBTC) and the Centre of Expertise Leisure, Tourism & Hospitality (CELTH).

Authors

Diana Korteweg Maris (CELTH / HZ Kenniscentrum Kusttoerisme)
Evelien Visser-Jonker (NBTC)
Evelot Westerink-Duijzer (CELTH / HZ University of Applied Sciences)
Hans Hoekstra (NBTC)

Photography

Cover: Marvin Kuhr
Backcover: Sjoerd Bracke & Cuno de Bruin

Graphic design

Kees Hoendervangers, www.dtp-plus.nl

NBTC
Nederlands Bureau voor Toerisme & Congressen

HZ UNIVERSITY OF APPLIED SCIENCES

CELTH Centre of Expertise leisure, tourism & hospitality

Centraal Bureau voor de Statistiek

Table of contents

Management summary 4

1. Introduction 6
   1.1 Explanation Statistics Accommodations 6
   1.2 Regionalising data 7
   1.3 Reading guide 7

2. Different regional levels 8
   2.1 Current statistics 8
   2.2 Inventorying tourist areas 8
   2.3 Initial exploration for regionalisation 9

3. Methodology 11
   3.1 CBS method: direct estimator 11
      3.1.1 Analysis strata 11
      3.1.2 Aggregation of strata 11
      3.1.3 Implementation 12
   3.2 Reliability of statistics 12
   3.3 Alternative method: small domain estimators 13
   3.4 Chosen method for regionalising tourism data 14
      3.4.1 Proposed method 14
      3.4.2 Use of CBS data 14
   3.5 Conclusion 15

4. Results 16
   4.1 Pilot areas 16
      4.1.1 Results Zeeland 17
      4.1.2 Results Drenthe 18
   4.2 All of the Netherlands 19
   4.3 Effect of sample size 21

5. Conclusion and discussion 24
   5.1 Conclusion 24
   5.2 Discussion 25

Annex A – Glossary 26
Annex B – Tourist areas 28
Annex C – CBS methodology 30
Annex D – Margins 33
Annex E – Implementation 36

Management summary

To achieve good policy-making in the tourism-recreation domain, data are needed. There is a great need for data at regional and local level on the number of guests and overnight stays in accommodations. CBS provides insight into the number of guests and overnight stays in the Netherlands through its Statistics Accommodations (SLA). They do this standard at a national level, at a provincial level, for the five major cities and for tourist areas. The question is: to what extent it is possible to further regionalise the CBS Statistics Accommodations data to municipalities or tourist-relevant regions (other than the CBS classification of tourist areas). The DDL project 'Regionalising tourism statistics' explored the possibilities.

First, the underlying methodology of the Statistics Accommodations (SLA) was thoroughly reviewed. The SLA is based on a sample among accommodation establishments, where the responses of participating establishments are being scaled up to the total population of accommodation establishments present in an area. This method of scaling up should also be used when compiling figures for smaller areas and/or municipalities

The dataset for 2023 was then accurately mapped via CBS' microdata environment. For each municipality and tourism-relevant region, the number of respondents in the dataset (sample) and how this compares to the total population were examined. It was thus checked whether the statistics may be published according to CBS' output guidelines. Based on this, it was found that the vast majority of COROP areas and tourism-relevant areas meet the CBS output guidelines and have sufficient respondents. At the municipal level, just under 20% of municipalities have sufficient respondents. Therefore, there seemed sufficient opportunities to continue the survey.

We then looked at the data more substantively. This is because reliable statistics depend not only on the number of respondents, but also on the spread in the response rate. Together, this determines the reliability margin of the statistics. If an area has a very similar offer (all similar accommodation, same type, same characteristics), then it is expected that it does not require such high respondent numbers to arrive at low margins. After all, presumably the spread in the response rate is already low then too. But if there are very different accommodations in an area, then the dispersion in the responses is probably also high and, despite reasonable numbers of respondents, a total figure of occupancy cannot be given with sufficient reliability.

It is common in statistics to work with a 95% relative confidence margin of 10%. CBS also uses this margin in its Statistics Accommodations. Therefore, we checked for which municipalities and tourist regions there are enough respondents while the results meet this reliability margin. The example below shows how a relative reliability margin of 10% should be interpreted.

Example: Suppose the number of overnight stays in hotels in 2023 in province X is estimated at 17,000. Practically, this means that with 95% certainty, the actual number of overnight stays in hotels in province X will be between 15,300 and 18,700 in the said period. These limits are equal to 10% below and 10% above the estimated number of 17,000.

Looking at the confidence margins of the statistics, it was found that only one in eight COROP areas meets the desired margin of under 10%. For tourist-relevant areas and municipalities, this is even much less. Thus, although the number of respondents for some large municipalities and tourist-relevant areas seems to be sufficient, the margins are still too high to calculate reliable statistics. We must therefore conclude that with the current data it is not possible to present reliable statistics at the municipal level. Even merging municipalities into tourism-relevant areas hardly offers any additional possibilities.

This exploration also had an unintended side effect. By mapping the margins on tourism statistics for different regional levels, it became clear that even at COROP level, margins can sometimes be quite large. Therefore, the desirability of continuing to publish these COROP figures is still being explored.

Finally, the study looked at how large the sample would have to be to do meet the desired reliability margin of up to 10%. To achieve this for all COROP regions, the sample would have to grow from 2,832 now to more than 6,000 participating accommodations. Such an expansion is not to be expected.

Unfortunately, therefore, it now appears that the SLA does not offer opportunities to compile more regional tourism statistics. To meet data needs at regional and local levels, other options will have to be explored.

1 Introduction

To achieve good policy-making in the tourism-recreation domain, basic data are needed. These basic data include figures on the number of guests and overnight stays in accommodation facilities. The Statistiek Logiesaccommodaties (or Statistics Accommodations in English) (SLA) of Statistics Netherlands (CBS) is an important data source for these figures and is accessed via CBS' open data platform Statline. This platform displays data for the Netherlands as a whole and data at provincial level. However, tourism-recreation policy is primarily shaped at regional and local level. There is therefore a great need for data at a regional and a local level.

1.1 Explanation Statistics Accommodations

The aim of the Statistics Accommodations (SLA) is to determine the number of guests and their overnight stays per type of Dutch accommodation. Guests are visitors who stay one or more consecutive nights in an accommodation. Overnight stays are all nights that guests spend in an accommodation. A group of 4 people who stay for 3 nights in an accommodation counts for 12 overnight stays. The number of guests and overnight stays is investigated by surveying accommodations. The population of accommodations consists of the following types of accommodations:

- Hotels, motels, guesthouses, apartments with hotel services, youth accommodations and bed & breakfasts with at least 5 sleeping places;
- Cottage sites with at least 10 sleeping places;
- Campsites with at least 4 tourist pitches. These are tourist pitches for short stays. Places that are rented for a longer period, such as annual pitches or seasonal pitches, are not included in the statistics;
- Group accommodations with at least 10 sleeping places

A detailed description of the terms used in the SLA can be found in Annex A. Accommodations for personal use, such as second homes or mobile homes on annual pitches at campsites, are not part of the population. Marinas and cruise ships are also not part of the population. Due to the lower limit, very small-scale providers are also not taken into account, such as private rental/rental of residential space.

A sample of approximately 30% is drawn annually from the total population of registered accommodations. This sample then receives a monthly questionnaire about the number of guests and overnight stays, specified by country of origin. The results are checked for outliers and any deviations from the previous year at an individual level and per area. The results are then weighted for the total population, per province and for the whole of the Netherlands.

The SLA data are available as open data via the CBS data portal Statline, for the Netherlands as a whole, for provinces and for grouped tourist areas. The available time series starts from 2012, with figures per month, per quarter and per year. The data are published monthly, usually about 2 months after the end of the reporting month. The SLA data correspond to the guidelines of the European statistical organization Eurostat and are input for European statistics on guests and overnight stays.

1.2 Regionalising data

As mentioned earlier, there is a great need for data at regional and local level. In the Tourism Data Center, NBTC and CBS have already jointly ensured regionalisation of the data to the level of COROP areas, clusters of municipalities for the purpose of regional research. However, this area division does not match the information needs of parties in the tourism-recreational field, such as municipalities and Destination Management Organizations (DMOs). The desire is to have data at municipal level or for tourist areas. In order to explore to what extent the need for further regionalised statistics can be met, the DDL project '*Regionalising tourism statistics*' was launched at the beginning of 2024. The findings of this project are presented in this report.

1.3 Reading guide

The remainder of this report is structured as follows. We start in Chapter 2 with a discussion of the different regional levels: we look at what data are currently available and explore the possibilities of regionalisation. Then, in Chapter 3 , we discuss the methodology used by CBS to calculate tourism statistics. We explain how this methodology can be applied to lower regional levels. Not only will we look at calculating the statistics themselves, but also at the precision of these statistics. We then present the results in Chapter 4 . The report concludes with a discussion and conclusion in Chapter 5.

2 Different regional levels

In this chapter, we explore the possibilities of creating data for municipalities or tourism-relevant areas from the Statistics Accommodations.

2.1 Current statistics

The starting point in this project is the Statistics Accommodations as made available by CBS via StatLine. This statistic provides insight into the number of guests and overnight stays per province.

These statistics are calculated from estimated occupancies, for which CBS uses a sample. This sample contains about a third of all opened accommodations. A different sample is drawn each year and the accommodations in this sample are asked to respond to a survey. In this survey, they indicate by month and country of origin how many guests have stayed at their accommodation and the corresponding number of overnight stays. The survey results are then used to estimate the number of guests and overnight stays at accommodations that are not in the sample or did not respond to the survey.

Besides the provincial tourism statistics from CBS, NBTC has been publishing statistics at the COROP level on their website since 2020. These figures are also based on the CBS sample. With figures at the COROP level, NBTC sometimes runs into issues of data reliability and too much risk of revealing data of individual businesses. With further regionalisation to municipal level, this will increase.

2.2 Inventorying tourist areas

In order to present reliable statistics at a regional level, enough accommodations and respondents are needed per region. We realise that this is unlikely to be the case for small municipalities in particular. Therefore, we not only consider the municipal level, but also look at the option of merging certain municipalities into tourism-relevant areas. These are tourist areas other than the tourist areas used by CBS when publishing statistics on StatLine. The wishes of provinces and DMOs regarding area divisions have been mapped out.

This query elicited responses from seven provinces: Friesland, Drenthe, Overijssel, Gelderland, Utrecht, North Brabant and Zeeland. The desired classification of tourism areas for these provinces can be found in Annex B. The extent to which this classification differs from the COROP classification varies per province. For Drenthe and Gelderland, for example, the overlap between COROP areas and tourist areas is quite large. For the province of Utrecht,

there is much difference, as the province consists of only one COROP area, but works with six tourist areas.

2.3 Initial exploration for regionalisation

To determine whether it is possible to display tourism statistics at the regional level, several questions need to be answered. These questions are shown in the diagram below and to answer these questions we use CBS microdata in an RA environment.

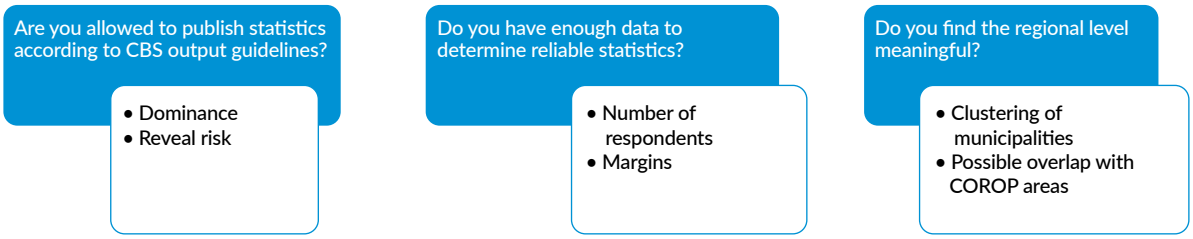


Figure 1 – Roadmap to achieve reliable and meaningful regional tourism statistics.

As shown in Figure1, the first question is whether it is permissible under CBS output guidelines to publish statistics at a certain level. These output guidelines are the criteria used by CBS and take into account disclosure risk and dominance of certain accommodations in a certain area. To answer this question, information on the number of accommodations per region and their capacity is sufficient. When it is allowed to publish statistics, the next question is whether it is also possible to determine reliable statistics. Here, the number of respondents plays an important role, as well as the margins on the calculated statistics. These margins provide a kind of bandwidth of uncertainty around the calculated occupancy. If the margins are too large, it is not possible to conclude anything about the statistics with sufficient certainty. The final question is whether the regional level of the statistics is sufficiently meaningful. This involves, for instance, whether the amalgamation of municipalities is useful for parties in the tourism and recreation field. The possible overlap with COROP areas may also determine whether the regional statistics have sufficient added value.

In our first exploration, we analyse the first two questions, looking only at the number of respondents for the second question. In following chapters, margins will be calculated. By accessing the microdata in the RA environment of CBS, the dataset of the year 2023 was analysed carefully in terms of population size and sample size at different area levels. The results are shown in Table 1

		Number meeting CBS output guidelines (dominance and disclosure)		Number with at least 10 respondents	
Regions	**Number**	**January**	**July**	**January**	**July**
Provinces*	13	13	13	13	13
COROP areas	40	40	40	36	38
Tourist areas	32	32	32	28	31
Municipalities	342	155	209	40	80

Table 1 – Analysis of the 2023 population and sample at different regional levels. Here we look at the total level and take all accommodation types together. (*) Hotels in Amsterdam are taken separately and separated from the province of North Holland. They thus form the thirteenth province, as it were.

In Table1 we see that all provinces meet the CBS output guidelines as well as having at least 10 respondents. This limit of 10 respondents was chosen in line with CBS' rules to only publish statistics based on at least 10 respondents. COROP areas and tourist areas almost always meet the criteria. When regionalising further to the municipal level, the picture changes. Only part of the municipalities meet the CBS output guidelines or the number of 10 respondents. We also see that clearly more municipalities comply in July than in January, which has to do with accommodations that are not open all year round. Nevertheless, almost half of the municipalities meet the CBS output guidelines and 12-23% of the municipalities also meet the number of at least 10 respondents

Based on these results, there is sufficient reason to continue the study. It seems possible to determine reliable tourism statistics at regional level for sufficiently large municipalities and tourist areas. To state with certainty whether sufficient reliability is guaranteed, we will have to calculate the confidence margins on the calculated statistics. We discuss this further in the next chapter.

3 Methodology

In this chapter, we discuss how tourism statistics can be calculated. To do so, we evaluate the method used by CBS and discuss what adjustments are needed to regionalise tourism statistics. We also discuss the reliability margins of the statistics.

3.1 CBS method: direct estimator

CBS estimates occupancy with a direct estimator based on sample data. The sampled accommodations report by month and country of origin the number of guests who stayed overnight with them and the number of overnight stays. Using this information, the number of guests and overnight stays is then also estimated for accommodations not in the sample.

3.1.1 Analysis strata

Analytical strata are used to calculate the direct estimator. All accommodations are assigned to a particular stratum (group) based on the accommodation type, size class, region, month and country of origin being reported on. To determine the estimated occupancy of an accommodation in a particular stratum, it is preferable to use only the data from the sample accommodation in that same stratum. This is only possible if a stratum has a sufficient number of respondents. In that case, an average occupancy is taken across all accommodations in that stratum and this average occupancy is applied to the accommodations not in the sample.

3.1.2 Aggregation of strata

When a stratum has too few respondents, the sample data for that stratum are not sufficiently reliable to make an estimate for the remaining accommodations. In that case, CBS uses a method where strata are aggregated. This method is described more precisely in Annex C.

When aggregating strata, certain priorities are applied to achieve reliable data based on a sufficient number of respondents. Aggregation by size class is the first preference. If that still leaves too few respondents, aggregation is done on the basis of region. Aggregation based on the other characteristics, i.e. accommodation type, month and country of origin of guests, is not done. This is probably because months and accommodation types are so different that merging would lead to loss of important information. Merging strata with different countries of origin does not yield additional respondents, as each accommodation in the sample reports data for each country of origin (even if the number of guests or overnight stays from that country of origin is 0).

3.1.3 Implementation

In implementing CBS' method, as described in Annex C, we ran into some ambiguities. The description of the method can sometimes be interpreted in several ways, making it unclear which implementation choices CBS itself made. Lack of clarity arose in particular around the questions of when analysis strata are merged and how this merging then takes place.

The project group submitted these implementation questions to CBS in July 2024. Unfortunately, due to circumstances, it took until December 2024 for CBS to provide clarity on their implementation. We have incorporated this response and describe in Annex E where the implementation differs from the documented methodology.

3.2 Reliability of statistics

Besides determining the estimated occupancy, we also calculate the precision of the obtained estimate. We do this using the relative 95% confidence margin. Annex D describes in detail how to calculate these margins. The relative confidence margin can be interpreted as follows.

> Suppose the number of overnight stays in hotels in January 2023 in province X is estimated at 17,000 and the 95% relative confidence margin is equal to 10%. This means that with 95% confidence, the actual number of overnight stays in hotels in province X is between 15,300 and 18,700 in the said period. These limits are equal to 10% below and 10% above the estimated number of 17,000, respectively.

If an estimate has a high margin, the interval in which the actual figure lies with 95% certainty is also large. This means that the actual figure may be much different from the estimate. Publishing estimates with a high margin is therefore not desirable, as it may confuse users and lead to wrong conclusions being drawn.

There could be several reasons for a high margin. First, the **number of respondents** plays an important role. When calculating regionalised statistics, there is a risk of increasing margins because statistics in each region are estimated on the basis of fewer respondents. The fewer respondents, the more difficult it can be to make a reliable estimate. Yet it is not only the number of respondents that is influential. Even more important is the **spread in the answers** given by respondents in the survey. In a municipality where accommodations are very similar (e.g. all medium-sized hotels in an urban area), a small number of respondents may be sufficient to make a good estimate of occupancy. When accommodations within a municipality are very different from each other (e.g. a combination of large and small hotels, campsites and/or group accommodations), many more respondents are needed to arrive at reliable occupancy figures. However, even when accommodations in a municipality are very similar but report very different occupancies, many respondents may still be needed. Due to large differences

in reported occupancies, it is difficult to determine whether the accommodations not in the sample are more similar to those with high or low occupancies. The overall figures for that municipality can then vary considerably and thus have a large margin.

So to make sure we report reliable figures, it is necessary to map margins properly and report figures only when we find margins low enough.

3.3 Alternative method: small domain estimators

During a consultation with CBS on July 16 2024, another alternative method for determining estimates at the municipal level was proposed. The use of small domain estimators is seen by CBS as the most promising direction for municipal statistics, as it uses as much data as possible and can reduce margins. These small domain estimators combine a direct estimator at the municipal level with a model-based estimate over a large region. When a municipality has enough respondents to produce a reliable estimate, the small domain estimator will lean more towards the direct estimator. For a municipality with too few respondents for a reliable estimate, the small domain estimator is mainly determined by the model-based estimate. By using small domain estimators, the margins can be reduced and a more reliable estimate can be determined at the municipality level.

To apply small domain estimators, a model needs to be created. This model predicts from data what the occupancy is in a given municipality. To build such a model, several choices have to be made. First of all, the level at which models are created must be determined. For instance, you can make a model for each month separately, but also for the whole year. You can make a model per province, but also for the whole of the Netherlands. You also have to choose which information is included in the model. Besides the available data from the sample, auxiliary information can be added that can improve predictions. In conclusion, working with small domain estimators requires some important choices for the model to be developed and possibly some experimentation with different models.

Small domain estimators have the advantage that they generally result in lower margins and hence more reliable statistics. How much the margins can be reduced varies from situation to situation, although CBS experts said the reduction in margins is often limited.

3.4 Chosen method for regionalising tourism data

In this section, we will discuss how we chose the method for determining regionalised data.

3.4.1 Proposed method

In the previous sections, two methods were discussed: the current CBS method via direct estimators and the alternative method via small domain estimators. After consultation with the project team, it was decided in July 2024 to opt for direct estimators within the DDL project. A key reason for this was that choosing direct estimators aligns with the method used by CBS, which also serves as the basis for the COROP figures published by NBTC. We are aware that the use of direct estimators may result in higher margins, which may require municipalities to be grouped to arrive at reliable statistics. However, we see it as an important goal of this project to explore the limits of what is feasible based on the current sample.

Another reason for this decision was that the remaining duration of the project and the budget still available were expected to be insufficient to achieve proper implementation of small domain estimators. Another factor was that the use of small domain estimators does not solve the problem of CBS output guidelines, which would still prevent from releasing statistics for many municipalities, no matter how small the margins.

3.4.2 Use of CBS data

To calculate direct estimators for tourism statistics, we need to determine how to use CBS microdata. These microdata contain for all respondents their reported occupancy, so estimates for these accommodations are not necessary. In addition, the microdata contain estimated occupancies for the remaining accommodations, both in terms of guests and overnight stays. Using estimated occupancy, tourism statistics can be calculated directly for non-respondents.

In Section 3.1, we explained how occupancy is estimated using the CBS method. Essentially, a weighted average is taken across comparable accommodation in a given region. CBS calculates two versions of estimated occupancy, with the regional level of estimated occupancy differing. There are estimates for each accommodation at provincial (PROV) level and at tourist area (TG) level[1]. This gives us the following opportunities to calculate data at a lower regional level.

1 The classification of tourist areas used by CBS is not the same as the classification of tourist areas in Annex B. There is a total of 17 tourist areas, while the number of tourist areas is larger.

Level at which estimated occu-pancy is calculated	Level at which data are published	Comment
PROV	PROV	Provincial statistics on StatLine.
PROV	COROP	NBTC's COROP statistics.
TG	COROP	A possible alternative to the NBTC's COROP statistics.
COROP*	COROP	A possible alternative to the NBTC's COROP statistics.
PROV	MUN	Opportunity in DDL project
TG	MUN	Opportunity in DDL project
MUN*	MUN	Opportunity in DDL project

Table 2 – Overview of possible combinations of the regional level of the estimate and the level at which the data is published. (*) Estimated occupancy at this level is not provided by CBS and has to be calculated itself.

Table 1 shows roughly two ways of calculating regional statistics. First, it is possible to directly use CBS' estimated occupancy at the province or tourist area level. Second, it is possible to determine your own estimated occupancy at a certain regional level. The first method is easy to apply, but has the major disadvantage that large differences within a province or tourist area will not show up in municipal statistics. After all, it is assumed that occupancy is the same on average per province or tourist area. The second method requires more complex calculations based on CBS survey data, but gives a more reliable picture of the situation per municipality. Another possible disadvantage of the second approach is that there are not enough respondents for all municipalities to arrive at reliable estimates. As a result, in some cases estimated occupancies at a higher regional level (COROP or province) will still have to be used.

After consulting the project team, we opt for the second approach, calculating the estimated occupancies ourselves at the municipal level according to the CBS method described in Annex C. In this way, the municipal situation is reflected as closely as possible and the statistics remain in line in terms of method with those at higher regional levels. It is worth noting that when we refer to 'municipal level' above, we mean a regional level that is more detailed than COROP level, but does not necessarily correspond to actual municipalities. We keep open the possibility of merging certain municipalities into tourism-relevant areas if municipalities themselves have too few respondents or municipal statistics lead to margins that are too high.

3.5 Conclusion

In this chapter, we have proposed a methodology for determining tourism statistics at the regional level. We opt for direct estimators and merge strata where necessary using the CBS method. When a region has too few respondents, we will aggregate strata at a higher regional level. We publish statistics only if the corresponding margins are low enough. That way, we avoid releasing unreliable information based on too few respondents, for example. In principle, the chosen method is applicable to the whole of the Netherlands.

4 Results

In the project plan the choice was made to carry out an initial analysis on a number of pilot areas. We have chosen the provinces of Zeeland and Drenthe as pilot areas. In Section 4.1 we will discuss the results from these pilot areas. Following this, in Section 4.2 we will give the results for the whole of the Netherlands. We conclude this chapter with a discussion of the effect of sample size on the reliability of the statistics in Section 4.3

The results in this chapter are based on analyses in CBS' microdata environment. Since CBS rules allow only statistics based on at least 10 respondents to be displayed, some tables will lack data for regions where this criterion is not met.

4.1 Pilot areas

For the pilot areas of Zeeland and Drenthe, we examine the following research questions at four regional levels: provincial, COROP area, tourist area and municipal.

1. Does the region meet CBS output guidelines? That is, are there at least 10 accommodations in the region and is the capacity of the largest accommodation less than 50% of the total capacity?
2. Is it possible to determine reliable statistics for this region? We look at the number of respondents and the margins of confidence of the statistics, considering a margin of confidence of up to 10% as acceptable.

4.1.1 Results Zeeland

Table 2 shows the answers to the above questions for the province of Zeeland, using the total number of guests in the month of July 2023 as a statistic.

		1. CBS output guidelines	2. Reliability	
Classification	**Region**	**Meets requirements?**	**# of respondents**	**Reliability margin**
Province	Zeeland	Yes	277	0.06
COROP	Zeeuws-Vlaanderen	Yes	57	0.23
	Overig Zeeland	Yes	220	0.06
Tourist area	Beveland & Tholen	Yes	27	0.38
	Schouwen-Duiveland	Yes	62	0.26
	Walcheren	Yes	131	0.20
	Zeeuws-Vlaanderen	Yes	57	0.23
Municipality	Borsele*			
	Goes*			
	Hulst*			
	Kapelle*			
	Middelburg	Yes	14	0.27
	Noord-Beveland	Yes	13	0.40
	Reimerswaal*			
	Schouwen-Duiveland	Yes	62	0.26
	Sluis	Yes	46	0.50
	Terneuzen*			
	Tholen*			
	Veere	Yes	106	0.26
		Yes	11	0.57

Table 3 – Overview of the regions in Zeeland indicating whether they meet CBS output guidelines and whether reliable statistics can be displayed. (*) These municipalities have fewer than 10 respondents, so no figures can be released.

From the above table, we can see that the CBS output guidelines have been met at almost all regional levels. At the provincial level, reliability is also guaranteed, given the low margins. However, these margins increase rapidly as we regionalise further. Already at the level of COROP areas and tourist areas, we see margins that are quite high, presumably due to the limited number of respondents per area. At the municipal level, this is even more pronounced. Only a limited number of municipalities have at least 10 respondents, allowing figures to be released. In addition, the margins for none of the tourist areas or municipalities are low enough to arrive at reliable statistics, assuming an acceptable margin of up to 10%.

We note that the results are based on total figures for the month of July 2023. For other months, such as January for example, these figures may look different and the number of respondents outside the summer season is likely to be even lower. The statistics are based on total figures and are therefore not broken down by accommodation type or origin of

guests. Further disaggregation may, on the one hand, result in lower margins by placing similar accommodations together. At the same time, further disaggregation means that there are fewer respondents per subgroup, which in turn has an adverse effect on margins.

4.1.2 Results Drenthe

We are also conducting a similar analysis for the province of Drenthe.

		1. CBS output guidelines	2. Reliability	
Classification	**Region**	**Meets requirements?**	**# of respondents**	**Reliability margin**
Province	Drenthe	Yes	162	0.06
COROP	Noord-Drenthe	Yes	68	0.15
	Zuidoost-Drenthe	Yes	42	0.17
Tourist area	Zuidwest-Drenthe	Yes	52	0.34
	Hondsrug	Yes	67	0.09
	Zuidwest Drenthe	Yes	52	0.34
	Midden-Drenthe	Yes	18	0.42
Municipality	Kop van Drenthe	Yes	25	0.33
	Aa en Hunze	Yes	25	0.25
	Assen*			
	Borger-Odoorn	Yes	19	0.21
	Coevorden	Yes	13	0.23
	De Wolden	Yes	44	0.33
	Emmen	Yes	10	0.52
	Hoogeveen*			
	Meppel*			
	Midden-Drenthe	Yes	18	0.42
	Noordenveld	Yes	12	0.93
	Tynaarlo*			
	Westerveld*			

Table 4 – Overview of the regions in Drenthe indicating whether they meet CBS output guidelines and whether reliable statistics can be displayed. (*) These municipalities have fewer than 10 respondents, so no figures can be released.

Drenthe province's results are similar to those of Zeeland. Reliable figures with low margins can almost only be guaranteed at the provincial level. Besides this general conclusion, there is another striking result. Namely, we see a low margin in the tourist area Hondsrug. This area consists of the COROP area Zuidoost-Drenthe, supplemented by one municipality from the COROP area Noord-Drenthe (Aa and Hunze). Apparently, the occupancy in Aa and Hunze is similar to that in the COROP area Zuidoost-Drenthe and adding this extra municipality provides enough extra respondents to arrive at a lower margin. This result shows that grouping similar municipalities can help to find lower margins. Similar results can also occur if results are broken down further, by reporting separate statistics for hotels, for example. Still, this is not necessarily the solution to the problem of high margins. Although accommodations

that are very similar (in a certain region or of a certain accommodation type) can be selected, there are also the remaining accommodations. For these accommodations, margins are not always lower. We also see this in Table 4. The COROP area Noord-Drenthe is divided between the tourist areas Hondsrug, Midden-Drenthe and Kop van Drenthe. Only Hondsrug has a low margin. On the contrary, the other tourist areas have higher margins than at COROP level.

4.2 All of the Netherlands

We then extend our results to the whole of the Netherlands. As we specifically focus on regionalising tourism statistics in this study, we only present the results here at the regional levels of COROP, tourist areas and municipalities. As the number of regions can be quite large, especially when looking at the municipal level, we limit ourselves in displaying the results. We choose to include areas in the results only if the margins on the target variable are at most 20%. As the target variable, we again use the total number of guests.

	January 2023			July 2023		
COROP	**Meets require-ments?**	**# of respondents**	**Margin**	**Meets require-ments?**	**# of respondents**	**Margin**
Noord-Drenthe				Yes	68	0.15
Zuidoost-Drenthe	Yes	19	0.14	Yes	42	0.17
Twente				Yes	121	0.15
Veluwe	Yes	106	0.16	Yes	181	0.10
Achterhoek				Yes	111	0.15
Utrecht	Yes	85	0.07			
Kop van Noord-Holland				Yes	112	0.20
Zaanstreek				Yes	10	0.12
Groot-Amsterdam	Yes	260	0.04	Yes	269	0.04
Agglomeratie Leiden en Bollenstreek				Yes	54	0.20
Agglomeratie 's Gravenhage	Yes	51	0.15	Yes	59	0.11
Groot-Rijnmond	Yes	79	0.15	Yes	98	0.09
Overig Zeeland	Yes	90	0.10	Yes	220	0.06
Zuidoost-Noord-Brabant	Yes	53	0.19	Yes	112	0.19
Noord-Limburg				Yes	74	0.14
Zuid-Limburg	Yes	111	0.12	Yes	159	0.07
Flevoland	Yes	27	0.20	Yes	44	0.18

Table 5 – Overview of COROP areas showing whether they meet CBS output guidelines and the corresponding measures for determining reliable statistics. Only areas with a margin of up to 20% in either period are included in the table.

The table above shows that 17 of the 40 COROP areas have a relative confidence margin of no more than 20% in January or July. This means that more than half of the COROP areas do

not meet this and thus have higher margins. We see in Table 5 that only a limited number of COROP areas have a margin of up to 10%. These are three areas in January and five areas in July. In particular, it is COROP areas with many respondents that result in a lower margin. A single exception is Zaanstreek which, with 10 respondents, still has a relatively low margin in July. This could have two possible causes. On the one hand, the respondents in this area may be very similar, resulting in a low spread in occupancy. On the other hand, it could also be coincidental that the reported occupancy of respondents in this area for July 2023 varies little. Only when there is an identifiable explanation, for instance because accommodations in that area are very similar, can it be expected that similar low ranges be found for this area for the dates of other months and years as well.

We then also present results at the level of tourist areas

	January 2023			July 2023		
Tourist areas	**Meets requirements?**	**# of respondents**	**Margin**	**Meets requirements?**	**# of respondents**	**Margin**
Hondsrug	Yes	28	0.20	Yes	67	0.09
IJsseldelta				Yes	15	0.18
Twente				Yes	91	0.20
Veluwe	Yes	110	0.16	Yes	188	0.10
Achterhoek				Yes	104	0.16
Utrecht	Yes	22	0.17	Yes	25	0.10
Walcheren				Yes	131	0.20
Midden-Brabant				Yes	50	0.17
Zuid-Oost Brabant	Yes	53	0.19	Yes	112	0.19
Waddengebied				Yes	177	0.16

Table 6 – Overview of tourist areas, indicating whether they meet CBS output guidelines and the corresponding measures for determining reliable statistics. Only areas with a margin of up to 20% in either period are included in the table.

The results from Table 6 are similar to the results at COROP level. The table above includes 10 tourist areas. This is less than a third of the 32 tourist areas, meaning that those not shown have a margin above 20%. Looking more closely at the table, we see that almost no tourist area has a margin below 10%. So here we see even more strongly than with the COROP areas that the vast majority of tourist areas have too high a margin to produce reliable results.

Finally, let us look at the results at the municipal level:

	January 2023			July 2023		
Municipalities	**Meets requirements?**	**# of respondents**	**Margin**	**Meets requirements?**	**# of respondents**	**Margin**
Ameland				Yes	32	0.16
Amstelveen	Yes	21	0.19			
Amsterdam	Yes	206	0.04	Yes	210	0.05
Arnhem				Yes	18	0.15
Haarlem				Yes	13	0.16
Haarlemmermeer	Yes	27	0.11			
Loon op Zand				Yes	15	0.17
Maastricht	Yes	27	0.14	Yes	30	0.19
Rotterdam	Yes	41	0.20	Yes	47	0.19
's Gravenhage	Yes	41	0.12	Yes	41	0.14
Texel				Yes	49	0.13
Utrecht	Yes	22	0.17	Yes	25	0.10
Voorst	Yes	21	0.19			
Zandvoort				Yes	16	0.17

Table 7 – Overview of municipalities indicating whether they meet the CBS output guidelines and the corresponding measures for determining reliable statistics. Only municipalities that meet the CBS output guidelines and have a margin of at most 20% in either period are included in the table.

The table above shows 14 municipalities, five of which are large cities (Utrecht, Amsterdam, Rotterdam, The Hague and Maastricht). Only the municipalities of Utrecht and Amsterdam have a reliability margin of up to 10% in January or July, which we consider acceptable. This is so limited that we must conclude that it is not possible to calculate reliable tourism statistics at the municipal level based on the current sample.

4.3 Effect of sample size

In previous sections, we have seen that the margins on the statistics tend to increase as figures are calculated at a lower regional level. This is partly explained by the decreasing number of respondents per region. It is therefore interesting to investigate whether increasing the sample size leads to margins of an acceptable level. We analyse in this section what increase in sample size is needed to achieve margins of 5% or 10%. Here, we use an approximation to broadly map the effect of sample size. [2]

2 Here, we assume the formulas for margins and variance as can be found in Annex D. Under the assumptions that the (1) sample variance does not change when increasing the sample and (2) the average size of accommodations in the sample and population are equal to each other, the required sample size m can be approximated as follows: $m = \frac{n}{f+(1-f)\frac{n}{N}}$. Here, n and N are the size of the current sample and population, respectively, and the factor f is equal to the square of the desired margin divided by the current margin.

		Total number of guests		Total number of overnight stays		Both statistics	
Regional level	**Current sample**	**Margins ≤ 10%**	**Margins ≤ 5%**	**Margins ≤ 10%**	**Margins ≤ 5%**	**Margins ≤ 10%**	**Marges ≤ 5%**
Provinces	2,832	3,353	4,120	3,809	5,441	3,809	5,441
COROP areas	2,817*	5,394	6,972	5,690	7,440	5,999	7,640
Tourist areas	1,737**	3,854	5,028	4,092	5,128	4,234	5,212

Table 8 – Number of respondents needed to get the 95% confidence margins small enough for two statistics, assuming the July 2023 sample. (*) This number is lower than at the provincial level because two COROP areas were not included, due to too few respondents to release data on them. (**) The tourist areas do not include all of the Netherlands and one tourist area had too few respondents to release data on.

The current sample size in July 2023 is 2,832 respondents out of a total of 9,058 opened accommodations. At the provincial level, margins are often already low. The sample only needs to increase by 12% to achieve a 95% confidence margin of at most 10% on both statistics in all provinces. To have a margin of at most 5% on both statistics, a larger sample is needed and the number of respondents will have to increase by 92%.

At the COROP level, we are missing data from two COROP areas that cannot be included due to too low a number of respondents. The figures from Table 8 for the COROP areas are therefore a lower limit, although we can assume that the percentage increase needed for the two missing areas is similar to the other COROP areas. To achieve margins of up to 10% at COROP level for both statistics, the number of respondents will have to increase to 6,000. This represents an increase of over 110%, i.e. more than a doubling. If the desired margin is equal to a maximum of 5%, a substantially larger increase in the sample is needed. For this to be achieved for both statistics at the COROP level, the sample size will have to be over 2.7 times larger.

For tourist areas, the size of the current sample is smaller as we did not receive input from all provinces regarding their current or preferred classification of tourist areas. Thus, these areas do not cover the whole of the Netherlands. However, also for these areas, we see that the sample must be 2.4 or 3 times larger, respectively, to arrive at relative confidence margins of at most 10% or 5%. Note that the current sample contains about a third of all opened accommodations. If the sample must be 2 to 3 times larger, this means that the sample must contain almost all opened accommodations. Given the previous results of margins at the municipal level, the pattern of increasing desired sample size will continue at the municipal level to achieve acceptable margins.

Finally, we note that in analysing the data from Table 8 , we have not taken non-response into account. If some of the accommodations in the sample do not respond to the CBS survey, an even larger sample will be needed to achieve the desired margins.

In conclusion, the huge expansions needed in sample size to arrive at reliable figures at the level of COROP areas or tourist areas do not seem feasible. Such expansions involve high costs. In addition, regionalised figures are not necessary for CBS, as they already comply with Eurostat obligations by publishing figures at the national and provincial level. Thus, increasing the sample to guarantee reliability for all areas at a given level is unlikely. However, it may be possible to increase the sample to a limited extent so that acceptable margins can be achieved for more areas than is currently the case

5 Conclusion and discussion

5.1 Conclusion

This report examines whether it is possible to regionalise tourism statistics. The starting point is the Statistics Accommodations currently made available at provincial level by CBS and at COROP level by NBTC. Because of the desire for more regional statistics, we are investigating whether these statistics can also be published at municipal level. In doing so, we include the option of merging municipalities into tourism-relevant areas if reliable figures at the municipal level are not possible.

It is only possible to publish reliable figures if the following conditions are met. First of all, it must be allowed by the CBS output guidelines to publish statistics. In addition, it must be possible to calculate reliable statistics. Here, the number of respondents plays an important role and the margins on the calculated statistics. These margins provide a kind of bandwidth of uncertainty around the calculated occupancy.

In terms of method, we choose direct estimators in our study and thus follow the CBS method. The chosen method is implemented in SPSS and is in principle applicable to the whole of the Netherlands.

Initial results show that the vast majority of provinces, COROP areas and tourist areas meet CBS output guidelines and seem to have sufficient respondents. At the municipal level, this is not always true, but still just under 20% of municipalities seem to have sufficient respondents. However, looking at the margins on the statistics, we see that even at the level of COROP areas, reliable statistics cannot always be guaranteed. This effect is amplified at the municipal level. The proportion of COROP areas that have a reliability margin of up to 10% is around 12%. For tourist areas, there is only one area and at the municipal level we find only two municipalities in the whole of the Netherlands that meet this. Thus, although the number of respondents for some large municipalities and tourist areas seems quite large, the margins are still too high to calculate reliable statistics. We must therefore conclude that based on the current sample, it is not possible to present reliable statistics at the municipal level. Even merging municipalities into tourism-relevant areas hardly offers any additional possibilities. Only when these areas are almost similar in composition and size to COROP areas, reliable figures can occasionally be determined.

5.2 Discussion

Exploring the possibility of regionalising tourism data has had an unintended side effect. By mapping the margins on tourism statistics for different regional levels, it becomes clear that even at COROP level, margins can sometimes be quite large. It is therefore recommended to examine whether it is still desirable to continue releasing these COROP figures without including an uncertainty margin.

Further research could also focus on possibilities of achieving reliable figures at a more detailed regional level than COROP level through other methods. For this purpose, it could be investigated what can be achieved with small domain estimators. This research should preferably be done in close cooperation with CBS, as they have extensive expertise in this field.

Finally, our results on the required sample size also offer starting points for follow-up steps. If there are opportunities to increase the sample size to a limited extent, it may well become possible to present reliable statistics for more municipalities or tourism-relevant areas than is the case based on the current sample.

Annex A – Glossary

Below is an overview of all the terms used in the Statistiek Logiesaccommodaties.

Guests: Visitors who stay one or more nights in an accommodation. A guest who stays in an accommodation for more than two consecutive months is considered a regular guest and does not count towards the statistics. Asylum seekers and seasonal workers are not counted as guests, even if they stay in an accommodation for less than two consecutive months. Each month, the guests who left in that month are recorded, regardless of the month in which the guest arrived. It is possible that a person is counted as a guest two or more times in one month in the same accommodation or in different accommodations.

Hotel: An accommodation with sleeping places for lodging in predominantly single and double rooms on a per night basis, where separate meals, snacks and drinks can be provided to guests and passers-by. A guest is a person who stays overnight in the accommodation in question and a passer-by is someone who does not stay overnight. In addition, other services can be provided, such as reception, room and telephone service. The statistics apply a lower limit of at least 5 sleeping places for these accommodations.

Motel: A location along the highway with interconnected rooms with doors to a parking lot or common area for lodging to be booked per night. The statistics use a lower limit of at least 5 sleeping places for these accommodations.

Pension: An accommodation with sleeping places for lodging in predominantly single and double rooms, where individual meals, small food items and drinks can be provided to guests but not to passers-by. The statistics apply a lower limit of at least 5 sleeping places for these accommodations.

Apartment with hotel services: Apartment which is being kept clean and in which the beds are made during the stay. The statistics use a lower limit of at least 5 sleeping places for these accommodations.

Youth accommodation: Youth hotel and youth hostel. A youth hotel is a hotel for mainly young guests with sleeping facilities that one does not have to share with 'strangers'. A youth hostel is an accommodation for mainly young guests with sleeping facilities in rooms and/or halls that one may have to share with 'strangers'. The statistics use a lower limit of at least 5 sleeping places for these accommodations.

Bed & breakfast: Private home where one can stay overnight and have breakfast. The statistics use a lower limit of at least 5 sleeping places for these accommodations.

Campsite: A terrain or part of a terrain with tourist sleeping places, where one can spend the night in tents, touring caravans, campers, tent houses or hikers' cabins. The statistics use a lower limit of at least 4 tourist pitches for these accommodations.

Cottage area: A site with a number of summer cottages, mobile homes, (holiday) bungalows or (holiday) apartments, which are mainly available for rent by the operator or manager of the complex. Apartments that are rented with hotel services are not considered as a cottage complex but as an (apartment) hotel. Apartments without hotel services, which are often part of a larger building, are considered as a bungalow or summer cottage. The statistics apply a lower limit of at least 10 sleeping places for these accommodations.

Group accommodation: Accommodation with lodging provision predominantly to persons in groups (not families) with sleeping facilities in rooms, halls, cottages, tent houses, apartments and/or tents that guests may have to share with strangers. The statistics apply a lower limit of at least 10 sleeping places for these accommodations. Group accommodations are understood to mean:

- camping farms
- (children's) holiday homes
- camp houses/scouting houses
- 'friends of nature' houses
- tent camps
- lodging facilities belonging to water sports centres or riding schools.

Annex B – Toeristische gebieden

Province	Tourist areas	Municipalities
Drenthe	Hondsrug	Aa en Hunze, Borger-Odoorn, Coevorden, Emmen
	Zuidwest Drenthe	Hoogeveen, Meppel, Westerveld, De Wolden
	Midden-Drenthe	Midden-Drenthe
	Kop van Drenthe	Assen, Noordenveld, Tynaarlo
Overijssel	Weerribben-Wieden	Steenwijkerland
	IJsseldelta	Kampen, Zwartewaterland, Zwolle
	Vechtdal	Dalfsen, Hardenberg, Ommen, Staphorst
	Salland	Deventer, Olst-Wijhe, Raalte
	Sallandse Heuvelrug	Hellendoorn, Rijssen-Holten
	Twente	Almelo, Borne, Dinkelland, Enschede, Haaksbergen, Hengelo, Hof van Twente, Losser, Oldenzaal, Tubbergen, Twenterand, Wierden
Gelderland	Rivierenland	Buren, Culemborg, Druten, Maasdriel, Neder-Betuwe, Tiel, West-Betuwe, West Maas en Waal, Zaltbommel
	Arnhem-Nijmegen	Arnhem, Beuningen, Berg en Dal, Doesburg, Duiven, Heumen, Lingewaard, Nijmegen, Overbetuwe, Renkum, Rheden, Rozendaal, Westervoort, Wijchen, Zevenaar
	Achterhoek	Aalten, Berkelland, Bronckhorst, Doetinchem, Lochem, Montferland, Oost Gelre, Oude IJsselstreek, Winterswijk, Zutphen
	Veluwe	Apeldoorn, Barneveld, Brummen, Ede, Elburg, Epe, Ermelo, Harderwijk, Hattem, Heerde, Nijkerk, Nunspeet, Oldebroek, Putten, Scherpenzeel, Voorst, Wageningen
Utrecht	Amersfoort e.o.	Amersfoort, Bunschoten, Eemnes
	De Utrechtse Heuvelrug	Baarn, De Bilt, Leusden, Renswoude, Rhenen, Soest, Utrechtse Heuvelrug, Veenendaal, Woudenberg, Zeist
	Het Groene Hart	IJsselstein, Lopik, Montfoort, Nieuwegein, Oudewater, Vijfheerenlanden, Woerden
	Gooi & Vecht	De Ronde Venen, Stichtse Vecht
	Kromme Rijnstreek	Bunnik, Houten, Wijk bij Duurstede
	Utrecht (stad)	Utrecht
Noord-Brabant	Noord-Oost Brabant	Bernheze, Boekel, Boxtel, 's-Hertogenbosch, Land van Cuijk, Maashorst, Meierijstad, Oss, Sint-Michielsgestel, Vucht
	Noord-West Brabant	Altena, Drimmelen, Geertruidenberg, Oosterhout
	West-West Brabant	Bergen op Zoom, Halderberge, Moerdijk, Roosendaal, Steenbergen, Woensdrecht
	Zuid-Oost Brabant	Asten, Bergeijk, Best, Bladel, Cranendonck, Deurne, Eersel, Eindhoven, Geldrop-Mierlo, Gemert-Bakel, Heeze-Leende, Helmond, Laarbeek, Nuenen Gerwen en Nederwetten, Oirschot, Reusel - De Mierden, Someren, Son en Breugel, Valkenswaard, Veldhoven, Waalre

Province	Tourist areas	Municipalities
	Zuid-West Brabant	Alphen-Chaam, Baarle-Nassau, Breda, Etten-Leur, Rucphen*, Zundert
	Midden-Brabant	Dongen, Gilze en Rijen, Goirle, Heusden, Hilvarenbeek, Loon op Zand, Oisterwijk, Tilburg, Waalwijk
Zeeland	Schouwen-Duiveland	
	Walcheren	Middelburg, Veere, Vlissingen
	Beveland en Tholen	Borsele, Goes, Kapelle, Noord-Beveland, Reimerswaal, Tholen
	Zeeuws-Vlaanderen	Hulst, Sluis, Terneuzen

*** Rucphen is included in the tourism region of Zuid-West Brabant. In the information provided by the province of Noord-Brabant, it was not clear to which region the municipality of Rucphen belonged. Based on the DMO classification, it was decided to place Rucphen under Zuid-West Brabant.**

Annex C – CBS methodology

1 Elevation

Four accommodation types (hotels, campsites, cottage sites and group accommodation) are being distinguished, but Amsterdam and Other are being analysed separately within hotels. For the output, the results of these hotels are aggregated. A separate grossing up takes place for each of these five accommodation types.

For the Netherlands as a whole, individual markup weights w_{hi} are determined per combination of accommodation type, size class and month, i.e. per (analysis) stratum *h*. The markup factors are calculated as the quotient of the number of accommodations N_h open in that month and the number of respondents r_h. No distinction is made within the strata between chain, combi and other accommodations, i.e. $w_{hi} = w_h = N_h/r_h$ for all accommodations *i* from stratum *h*. Differences in inclusion probabilities are thus neglected (this was already done in the past for the chain accommodations, but not for the combi accommodations). This is because occupancy rates by country of origin differ little systematically between these three types of accommodation. We assume that the bias (degree of impurity) due to these aggregations for the estimates of numbers of guests and overnight stays is smaller than the additional variance that the differences in weights would otherwise produce. As a result, all accommodations in an analysis stratum 'accommodation type x region x month x size class' receive the same increment weight N_h/r_h krijgen. Only for small numbers of sample accommodations per stratum are the uplift factors used, due to aggregation of size classes (step 3 in section 1.2). Exceptionally, if a weight w_h is greater than 10, the weight is truncated to 10. Again, this only affects the results when size classes are merged [Yuri].

2 Numbers of guests and overnight stays by accommodation type x month x region

We give below the estimation method for the number of guests per analysis stratum accommodation type x size class x region x month. Region here can be the Netherlands region as well as provinces, COROP areas, etc.

Per analysis stratum: lodging type x size class x region x month (*h*):

y_{hi} =score on target variable *y* (number of guests or overnight stays by country of origin) of accommodation (respondent) *i* of stratum *h*;

x_{hi} = capacity of accommodation *i* of stratum *h* (per day);

D_h = number of days in the month at stratum *h*;

$R_{hi} = \frac{y_{hi}}{x_{hi}D_h}$ = occupancy in accommodation *i* of stratum *h*.

Per analysis stratum h, the following steps take place.

1 For each target variable *y*, the average occupancy per bed is estimated as the quotient of the number of guests or overnight stays and the capacity at the respondents.

$$\hat{R}_h = \frac{\sum_i w_{hi} y_{hi}}{\sum_i w_{hi} x_{hi}} \frac{1}{D_h}. \quad (1)$$

The accommodations are thus weighted in numerator and denominator with the uplift factors $w_{hi} = N_{(h)}/r_h$. This does not affect the outcome because within the analysis stratum no distinction is made between chain, combi and other accommodations, and thus all accommodations have the same weight $w_{hi} = w_h$. Thus, there is

$$\hat{R}_h = \frac{\sum_i w_{hi} y_{hi}}{\sum_i w_{hi} x_{hi}} \frac{1}{D_h} = \frac{\sum_i y_{hi}}{\sum_i x_{hi}} \frac{1}{D_h} \quad (2)$$

Only when size classes are merged, the last equals sign is dropped.

2 For non-respondents (non-surveyed and non-respondents), the estimates $\hat{R}_h$ are imputed as occupancy $\hat{R}_{hi}$. Then the numbers of guests and overnight stays are calculated as the product of the corresponding occupancy and capacity

$$\hat{y}_{hi} = \hat{R}_{hi} x_{hi} D_h. \quad (3)$$

Respondents retained their reported numbers of guests and overnight stays.

3 When an analysis stratum has fewer than 5 respondents, it is merged with the stratum with a higher or lower size class: GK 1 and 2 can be merged, and GK 3 and 4. The occupancy to be imputed for the unobserved accommodations is then calculated over the two size classes (within accommodation type x region x month), with the accommodations receiving as weights the markup factors calculated in section 5.1 for each original EO. In that case, formula (1) is applied instead of (2). However, when at least 40% of the accommodations are observed, the stratum is not aggregated.
If GK1+2 or GK3+4 has less than 5 respondents, all size classes are merged.

4 If within a region (by type x month) less than 8 accommodations are observed, the occupancy is taken from the whole of the Netherlands.

5 Results are summed by accommodation type x region x month across size classes h:

$$\hat{Y} = \sum_{st.pr.} y_{hi} + \sum_{pop\setminus st} \hat{y}_{hi} \,. \tag{4}$$

The first sum sign represents the numbers of overnight stays (guests) observed in the sample, and the second sum sign represents the imputed numbers of overnight stays (guests) for the unobserved accommodations. This formula provides the estimated total number of overnight stays (and guests) per month by accommodation type for each region (Netherlands, province, COROP etc.).

Quarterly and annual figures are derived from formula (4) by summation over the months.

The above procedure is also followed for the number of stars in hotels.

3 Aggregation across regions and correction factors

The provinces form the basis for the national figures. Totals for the Netherlands are calculated by summation across the 12 provinces. Amsterdam is hereby analysed separately as part of the Province of North Holland. Thus, after calculating the estimated province totals according to formula (4), the following follows

$$\hat{Y}_{Ned} = \sum_{prov} \hat{Y}_{prov} \tag{5}$$

The province total $\hat{Y}_{prov}$ (guests and overnight stays by country of origin) calculated according to (4) will usually differ from the sum of the estimated totals for the COROPs belonging to the province, $\sum_{corop \in prov} \hat{Y}_{corop}$. For consistency, a correction factor is calculated as the quotient of both, and the COROP totals $\hat{Y}_{corop}$ are multiplied by this:

$$\hat{\hat{Y}}_{corop} = \hat{Y}_{corop} (\hat{Y}_{prov} / \sum_{corop \in prov} \hat{Y}_{corop}) \,. \tag{6}$$

The COROP totals are thus calibrated to the province totals.

For most other characteristics, the categories cross provincial boundaries. The category totals are then calibrated to total Netherlands, i.e. to the sum across provinces.

Annex D – Margins

Written explanation by CBS, August 2024

Occupancy is used in the Statistics Accommodations when making estimates. The following discusses making estimates in such a scenario and determining the corresponding margins based on a quotient estimator. The theory described here is derived from that in Särndal et al (1992).

Important target variables for the survey of accommodation establishments are the population total per month for the number of guests and the number of overnight stays per accommodation type. In addition, it is common to distinguish between Dutch and non-Dutch guests (with possibly a further breakdown by country of origin of the guests). We denote these target variables with the letter Y, e.g. the number of overnight stays per accommodation type in the Netherlands. Let $\hat{y}$ be an estimate for Y. Y is estimated using various analysis strata, based on size class, accommodation type, region and month. Here, each analysis stratum uses a quotient estimator with capacity as the auxiliary variable. This is because there is a strong relationship between the number of overnight stays and the available capacity of an accommodation. Moreover, the quotient of the two variables seems to fluctuate less than the target variable itself, so the use of the quotient estimator has a positive impact on making the estimates.

In the remainder of this paper, stratum refers to the analysis stratum. Let h = 1, ... , H be the different strata and let U_h represent the population of accommodations in stratum h. Let ρ_h represent the set of responding accommodations from stratum h. Let n_h be the number of responding accommodations in stratum h and let N_h be the number of accommodations in the population in stratum h. Let y_{hk} be the value of target variable for (responding) accommodation k in stratum h. Let x_{hk} be the capacity (per day) of accommodation k in stratum h, let D_h hbe the number of days in the month belonging to the month corresponding to stratum h and let R_{hk} be the average occupancy per day at accommodation k in stratum h, i.e.

$$R_{hk} = \frac{y_{hk}}{x_{hk} D_h},$$

where y_{hk} is the number of overnight stays for respondent k on a monthly basis nd x_{hk} dis the capacity per day. Note that R_{hk} can only be determined for respondents. If $\hat{R}_h$ is used to note the estimate for average occupancy per day for stratum h , then

$$\hat{R}_h = \frac{\sum_{k \epsilon \rho_h} y_{hk} w_{hk}}{\sum_{k \epsilon \rho_h} x_{hk} D_h w_{hk}},$$

where w_{hk} is a boost factor with w_{hk} = min(N_h/n_h , 10). In general

$$\hat{R}_h = \frac{\sum_{k\in\rho_h} y_{hk} w_{hk}}{\sum_{k\in\rho_h} x_{hk} D_h w_{hk}} = \frac{\sum_{k\in\rho_h} y_{hk}}{\sum_{k\in\rho_h} x_{hk} D_h},$$

unless strata are aggregated, as when aggregating size classes. Then, for non-respondents, the estimated target variable $\hat{y}_{hk}$ is determined as follows

$$\hat{y}_{hk} = \hat{R}_h x_{hk} D_h.$$

Next, the population total for stratum h, $\hat{y}_h$, can be estimated as follows

$$\hat{y}_h = \sum_{k\in\rho_h} y_{hk} + \sum_{k\in U_h\setminus\rho_h} \hat{y}_{hk},$$

from which an estimator for Ysimply follows: $\hat{y} = \sum_{h=1}^{H} \hat{y}_h$.

It is further important to know the precision of the obtained estimator. The precision can be quantified through the relative 95% confidence margin $M(\hat{y})$. It can be calculated by:

$$M(\hat{y}) := \frac{1{,}96 \cdot \sqrt{Var(\hat{y})}}{\hat{y}}.$$

The margin is proportional to the root of the variance of the estimator. The smaller the variance, the smaller the margin and thus the more accurate the estimate. The variance $Var(\hat{y})$ depends on all values of the target variable in the population. It cannot be calculated because the value of the target variable is known only for the responding accommodations, and thus will have to be estimated from the response rate. To arrive at an estimate for the variance of the total across all strata, it is useful to be able to estimate the variance by stratum, $Var(\hat{y}_h)$. This can be approximated, for strata that are not pooled, by the following formula

$$\widehat{Var}(\hat{y}_h) = \left(\frac{\sum_{k\in U_h} x_{hk}/N_h}{\sum_{k\in\rho_h} x_{hk}/n_h}\right)^2 {N_h}^2 \frac{1 - {n_h}/{N_h}}{n_h} {S_{es_h}}^2,$$

see Särndal (1992), where

$${S_{es_h}}^2 = \frac{1}{n_h - 1} \sum_{k\in\rho_h} (y_{hk} - \hat{R}_h x_{hk})^2,$$

the sum of the square of the residuals. These two formulae change at the point when strata are combined. In such cases, the following approach is used for the estimated variance per combined stratum h

$$\widehat{Var}(\hat{y}_h) = \left(\frac{\sum_{k\in U_h} x_{hk}/N_h}{\sum_{k\in\rho_h} x_{hk}/n_h}\right)^2 {N_h}^2 \frac{1 - {n_h}/{N_h}}{n_h} {\tilde{S}_{es_h}}^2,$$

in which

$$\tilde{S}_{es_h}{}^2 = \frac{1}{n_h - 1} \sum_{k \in \rho_h} (w_{hk}(y_{hk} - \hat{R}_h x_{hk}))^2,$$

where the weights are used both to determine $\hat{R}_h$ and to sum the squared residuals.

Using the estimated variance of $\hat{y}_h$, the 95% confidence margin for each stratum can then be estimated and is equal to

$$\widehat{M}(\hat{y}_h) := \frac{1{,}96 \cdot \sqrt{\widehat{Var}(\hat{y}_h)}}{\hat{y}_h}.$$

e variance for the total across all strata combined is further $\widehat{Var}(\hat{y}) = \sum_{h=1}^{H} \widehat{Var}(\hat{y}_h)$

and the estimated margin for the total $\widehat{M}(\hat{y}) := \frac{1{,}96 \cdot \sqrt{\widehat{Var}(\hat{y})}}{\hat{y}}$.

References

- C.-E. Särndal, B. Swensson & J. Wretman (1992) Model assisted survey sampling. Springer Verlag, New York Inc. Springer series in Statistics.

Annex E – Implementation

As described in Section 3.1.3, CBS' methodology can sometimes be interpreted in multiple ways, making it unclear what implementation choices CBS itself has made. Lack of clarity arose in particular around the questions of *when* analysis strata are merged and *how* this merging is then done. These questions were submitted to CBS, after which some adjustments were made to our implementation. In this Annex, we explain our implementation, which is as close as possible to that of CBS, but does differ in certain respects from the method documented by CBS in Annex C.

Abnormal response

When a respondent's reported occupancy deviates too much from the other respondents, that response is considered a outlier . CBS uses the criterion that outliers deviate more than 5 standard deviations from the average occupancy for guests or more than 15 standard deviations for overnight stays. The microdata indicate which respondents are considered outliers. We include these responses of outliers in our implementation for their own occupancy too, but do not include them in determining the estimated occupancy for non-respondents.

Besides anomalous responses, it is also notable that in some strata all or many respondents report a response of 0. That is, they report having had no guests or overnight stays in that month. CBS indicates not to include this response in their implementation. We do choose to retain this response, because even a value of 0 contains information, for example because in a given month there were no guests from a rare country of origin. We assume that incorrectly reported values of 0 are filtered out in the definition of outliers. It would require more precise research to determine exactly when a reported value of 0 should be considered unusable.

Introducing terminology; partners and family

When an analysis stratum has too few respondents, it is sometimes merged with another analysis stratum. This considers the size class of the accommodations and distinguishes between '*large*' and '*small*' accommodations. We assume that for hotels, accommodations with size classes 5 and 6 are considered *large* and accommodations with size classes 2, 3 or 4 are considered *small*. For other accommodation types, we consider size class 4 as *large* and size classes 2 and 3 as *small*.

To define more precisely how we aggregate analysis strata, we introduce the following terminology. They h any stratum. We define the family strata of h as follows: all family strata of h are about the same region, lodging type, time period and country of origin as h, but not about the same size class. In addition to the family strata, we also introduce the *partner strata* of any stratum h in the following way. Partner strata of stratum h are family strata of h, but have a size class in the same category (large or small) as the size class of stratum h. Tabel A shows family relationships between size classes.

Accommodation type	Size class of stratum	Size classes of partner strata	Size classes of other family strata
Hotels	2	3, 4	5, 6
	3	2, 4	5, 6
	4	2, 3	5, 6
	5	6	2, 3, 4
	6	5	2, 3, 4
Other (campsites, cottages and group accommodation)	2	3	4
	3	2	4
	4	-	2,3

Table A – The partner and family relationships between the different size classes.

Merging strata

Using the defined partner strata and family strata, we can discuss how to merge strata when there are too few respondents. To do so, we use the decision structure below, which goes through each analysis stratum. If a question is answered 'yes', then continue the structure to the bottom right (green). If the answer is 'no', you go one step to the bottom left (orange).

Do the stratum, partner strata and all other family strata together have at least 9 respondents?		
Take over the occupation from a higher regional level.	Within the family, do both the stratum with its partner strata and the other family strata each together have at least 6 respondents?	
	Merge all analysis strata within the family.	Do both the stratum and partner strata have sufficient* respondents?
		Merge the stratum with the partner strata.
		The stratum does not need to be merged.

Table B – Decision tree for merging analysis strata.
* A stratum has sufficient respondents if either there are at least 6 respondents, or the respondents in the stratum cover at least 60% of the capacity in that stratum.

We note that the different limits in Table B are corrections to the methodology in Annex C. The definition of when a stratum has a sufficient number of respondents is also slightly modified.

Merging at an overarching level

The decision structure in Table B shows that at least 9 respondents in a family are needed to arrive at a reliable occupancy. If this is not the case, the occupancy is taken from a higher regional level. For the higher regional level, we take the national level in case we calculate statistics at the provincial level. For statistics at the other regional levels (COROP, tourist areas, municipalities), we take the provincial level as the higher regional level.

The margins must also be calculated differently in this case. Here, we take the sample standard deviation at the provincial level and impute it for all strata within the family in question. This results in the same sample standard deviation for all these strata, because in this case we are aggregating all size classes at the provincial level. We continue the regular calculation for the margins.

NBTC
Nederlands Bureau
voor Toerisme &
Congressen

HZ UNIVERSITY
OF APPLIED SCIENCES

CELTH
Centre of Expertise
leisure, tourism & hospitality

Centraal Bureau
voor de Statistiek