Adjusting for Cell Suppression in Commuting Trip Data

BY Christian Braathen, Inge Thorsen and Jan Ubøe

DISCUSSION PAPER

Institutt for foretaksøkonomi
Department of Business and Management Science

FOR 13/2022

ISSN: 2387-3000
December 2022

Adjusting for Cell Suppression in Commuting Trip Data

Christian Braathen , Inge Thorsen, and Jan Ubøe *

Abstract

Maximum entropy methods are used to infer the true trip-distribution matrix in cases where parts of the data are suppressed due to privacy concerns. Large proportions of the suppressed data are found to be inferred correctly when the marginal totals in the trip-distribution are known. Entropy-based approaches are further found to outperform a strategy of ignoring suppressed information in cases with suppressed marginal totals and/or a higher cut-off value of suppressing cell information. Our methods are demonstrated to reduce the systematic bias in estimates of the distance deterrence parameter to such small numbers that it is effectively zero, preventing potentially serious bias in estimates and predictions resulting from standard spatial interaction models. Another useful contribution is to identify what scenarios an entropy-maximization approach benefits from incorporating information on times series and/or information on distances in the transportation network.

1 Introduction

Privacy concerns explain why statistical agencies often introduce limitations in releasing data, of which cell suppression is the most commonly used approach for tabular data. This applies, for example, to tabular data on commuting trips between Norwegian census tracts. At this level of spatial interaction, data are generally suppressed for all combinations of census tracts with less than three commuters. This limitation in data is not introduced for more aggregated subdivisions of the geography, for instance, at the municipality level. Particularly for sparsely populated, rural areas, the cell suppression may involve a considerable number of commuters and potentially lead to seriously biased estimates, for example, on how variations in distance affect the travel demand. Data on commuting trips may not be very sensitive on their own. However, a few

*Corresponding author: Norwegian School of Economics, Helleveien 30, N-5045 Bergen, Norway, e-mail: jan.uboe@nhh.no, phone: 004755959978, fax: 004755959650

observations in a cell can be linked to information in other databases to be used to identify individuals and more sensitive attributes.

Based on synthetic data on commuting trips, we discuss several issues related to statistical disclosure limitations:

- Methods to reconstruct the true matrix, both in cases where we have observations just for a specific point in time and in cases where we are utilizing time-series information on commuting trips, as well as information on distances between specific origins and potential destinations.
- Is the statistical limitation in releasing data, like cell suppression, ignorable for analytical purposes, or does it influence the results substantially in terms of biased parameter estimates and predictions?
- Does the common practice of cell suppression represent a reasonable balance between privacy and accuracy?

Methods to limit disclosure of confidential data are frequently used, representing a challenge for empirical research. As stated in Abowd and Schmutte (2019), "The threats to privacy inherent in the "big data" era have affected the policies governing statistical agencies." According to Abowd and Schmutte (2016), economists "have not fully understood how pernicious suppression bias is." Carpenter et al. (2022) demonstrate how imperfect replication of suppressed cell information represents a source of measurement error that may increase standard errors and/or lead to biased parameter estimates in econometric analysis. Abowd and Schmutte (2016) recommend that more effort be made to develop other privacy-preserving methods and that a standard should be developed for citing data and for appropriate documentation and discussion of the relevant statistical disclosure limitation (SDL).

The primary motivation for this paper is threefold. First, we discuss how to correct adequately for ordinary cell suppression in problems focusing on the distribution of commuting trips between different towns in a geography. Second, we discuss the need for the statistical agency to tighten up the usual practice of cell suppression to preserve a sufficient level of confidentiality. Third, we discuss how statistical limitations in releasing data affect the results following a standard trip distribution model. Are the parameter estimates and the model performance sensitive to cell suppression procedures, or is suppression ignorable in studying commuting? Carpenter et al. (2022) provide a recent general discussion of how suppressed cell estimated data sets may bias statistical inference.

The analysis in this paper is conducted in terms of spatial interaction, represented by the journey-to-work. Still, we think that the central part of the analysis can be generalized and transferred to

other problems. The results are potentially helpful for researchers reflecting on how statistical limitations in releasing data affect their estimation and predictions and suggesting methods to deal with, for instance, cell suppression. The results are also potentially useful for statistical agencies in designing rules on how to suppress information for the public in preventing sensitive data from being disclosed.

Our results are not based on observations from an actual geography. Instead, data are generated from an agent-based approach. A population of interacting, utility-maximizing agents make labor and housing market decisions in a geography of 12 towns. From the initiation, the inhabitants are randomly assigned characteristics according to frequencies observed in Norwegian data. Hence, the synthetic population in our fictive geography mirrors the Norwegian population in terms of, for instance, age, getting married, the number of children, divorce rates, being retired, death rates, etc. An agent-based approach to providing data for the analysis has the advantage that different aspects concerning confidentiality and trip distributions can be systematically monitored and examined, for example, to avoid issues related to endogeneity and causality.

Section 2 provides a review of literature discussing different issues concerning statistical disclosure limitation (SDL). This involves the trade-off between privacy and accuracy in estimation and predictions, alternative methods for maintaining privacy in publicly released data, and methods to adjust for SDLs. Section 3 introduces the geography and explains how the synthetic data used for the analysis are generated from an agent-based approach and, as such, are not exposed to privacy concerns. Section 4 demonstrates how an entropy-maximizing approach successfully discloses suppressed information on commuting trips based on data stretching over the 20-year-long period we consider. As reported in Section 5, incorporating information on time series in an entropy-maximizing approach significantly discloses suppressed information. In contrast, distance information does not turn out to be relevant in the case with a cut-off suppression value of 3. In Section 6, we demonstrate that leaving out information on marginal totals does not lead to a substantial limitation on the possibility of disclosing suppressed information about commuting trips. The experiments further suggest that a procedure with entropy maximization has the potential of removing a substantial source of bias, also in cases with a higher cut-off value of cell suppression. As reported in Section 7, we also find that distances contribute significantly to better fitting in cases with sparse information on the pattern of suppressed information, while time series information no longer contributes. Section 8 demonstrates that ignoring the commuters represented by the suppressed cells causes biased estimates of the distance deterrence parameter in a standard doubly constrained gravity model. This bias is, to a large degree, eliminated when the suppressed information is disclosed by entropy-maximization. Finally, concluding remarks are provided in Section 9.

2 Statistical Disclosure Limitation; a Review of Relevant Literature

The increasing access to microdata has reinforced issues related to privacy and confidentiality. As pointed out by Matthews and Harel (2011), data releasing agencies must account for the fact that privacy is viewed as a fundamental human right by the United Nations. One issue is that some agents may choose to make malicious use of data, and another is that respondents may resist providing honest answers if they can be disclosed by outsiders (Matthews and Harel, 2011).

On the other hand, "more powerful computers and advances in algorithms such as machine learning have led to an explosion in the usefulness of data", see Jones and Tonetti (2020), who claim that the "economics of data" and machine learning is a rapidly growing field. Microdata provides the basis for sound empirical research, for example, in medicine, social issues, and economics. Hence, confidentiality concerns must be balanced against the utility of reaching new insight with potentially important policy implications. In an optimization context, the released data should minimize the loss of information necessary to preserve the privacy of the individuals in the database. Data publishers should be aware of this trade-off between confidentiality and potential usefulness in research. Abowd and Schmutte (2019), refer to literature proving that there is no free lunch, which means that any publishing of useful statistical summaries gives a loss of privacy. At the same time, suppressing information due to privacy concerns introduces a measurement error to the data. This may, in particular, influence the analysis in cases where different categories of workers and industries are introduced for a spatially disaggregated specification of a geography in a sparsely populated, rural area. Carpenter et al. (2022) discuss measurement errors in general and in county-level economic data in particular. They also demonstrate that the biases due to this kind of measurement error are severe for panel data estimators, and they point out that there is a trade-off between the bias that can be expected from using an aggregate representation of a variable and the bias resulting from more suppressed information if a more disaggregated approach is followed.

Concerning the potential utility of more accurate information, the literature also provides an interesting perspective based on a political-agency framework. Binswanger and Oechslin (2020) demonstrate that access to more data may prevent reform attempts and give incentives to preserve the status quo. The basic idea is that politicians are not always guided by the common interest but rather, for example, by re-election concerns. In such a scenario, imprecise and inadequate economic statistics can be used as a strategy of defense for reforms that proved to be failing. With better access to reliable data, the outcome may be to delay reforms, and more publicly available information may lead to under-experimentation (Binswanger and Oechslin (2020)).

2.1 Defining and Measuring Privacy and Accuracy

Consider first different concepts concerning privacy. Willenborg and De Waal (2012) distinguish between the risk of re-identification, which relates to the identification of an individual, and a predictive disclosure, which is about identifying values of a sensitive attribute for an individual. Abowd and Schmutte (2016) denote the first as an identity disclosure, while G. Duncan and Lambert (1989) distinguish between attribute disclosure and inferential disclosure. The former relates to obtaining reliable information due to linking attacks. At the same time, the latter occurs when reliable information is reached even without linking to observations from another database. The possibility of linking is related to trail disclosure, occurring when "an individual's location access pattern can be matched across the shared databases" Airoldi et al., 2011. G. Duncan and Lambert (1989) discuss population, or model, disclosure, which is about reaching confidential information about a population using a model based on the released microdata. In more recent literature (Abowd and Schmutte (2016)), this kind of disclosure is treated probabilistically, reflecting the ability that the published data can be used to identify individuals or attributes with a substantially higher probability. As pointed out by Abowd and Schmutte (2016), this applies, for instance, in approaches that use the randomized response to sensitive questions. This SDL method will be briefly discussed in Section 2.3.

Matthews and Harel (2011) provide a relatively detailed discussion of how privacy can be measured. One basic idea is to measure to what additional degree information from a database threatens the risk of re-identification and/or predictive disclosures. If the risk is high, this is an argument that the released microdata is limited to a few variables or that the data is massively modified before being released (Paass, 1988).

In addressing the measuring and definitions of privacy, both Shlomo (2018), Chetty and Friedman (2019), and Abowd and Schmutte (2019) are focusing on the concept of differential privacy, building on ideas from for example Dwork et al. (2006). This concept is based on the introduction of a parameter ε, which is measuring "the maximum difference in the log odds of observing any statistic across similar databases" (Abowd and Schmutte (2019), that are distinguished by whether confidential data are included or not. Hence, the privacy loss is limited by defining ε as an upper bound of the likelihood ratio (Chetty and Friedman, 2019), corresponding to a maximum risk accepted for a released statistic for preserving privacy. This parameter reflects the possibility that potential intruders can identify sensitive individual information from published statistics. Hence, privacy loss is also reflecting the availability of possible external information. In other words, the differential privacy is related to the loglikelihood ratio defined by the likelihood that the released parameter estimate stems from a specific dataset rather than a dataset that differs by one observation. If this likelihood ratio falls, for example, due to noise infusion, it will be more difficult to distinguish between two marginally different datasets. Hence, this will

increase the likelihood that privacy is preserved even if parameter estimates based on a small sample of individuals are released.

Abowd and Schmutte (2019) define the statistical accuracy "as the expected squared error between the published value and the value that would be published in the absence of privacy protection",

Many different methods to measure privacy was introduced in the 80s and the 90s; see, for instance, Matthews and Harel (2011) for a review. Sweeney (2002) introduced the measure k-anonymity, achieved if a specific combination of individual characteristics appears at least k times in a table. Grouping and/or cell suppression are techniques that can be introduced to achieve k-anonymity. A potential problem with this measure is that disclosures may find a place if there is a lack of diversity in the occurrence of sensitive attributes and if a potential intruder has some background information on specific individuals and/or attributes. The literature provides suggestions on overcoming such problems; see, for instance, Matthews and Harel (2011). They, in addition, discuss the presence of inferential disclosure, where information on some attributes, like gender, age, occupation, and region (explicit identifiers), can be combined with a more sensitive variable to build a regression model, reaching a predictive distribution of the values of the sensitive variable. As claimed by Abowd and Schmutte (2016), it is, from a probabilistic perspective, in general, impossible to release data without compromising confidentiality.

Heldal and Fosen (2001) also discuss issues related to inferential disclosure, starting by quoting Dalenius and Reiss (1982): "If the release of the statistics S makes it possible to determine a (microdata) value D more accurately than is possible without access to S, then a disclosure has taken place." This means that any publication of statistical data represents a disclosure, which can be formalized by a Bayesian approach where the observed S is used to update an apriori distribution into a more accurate aposteriori distribution $f_D(D|S)$. Heldal and Fosen (2001) further discusses uncertainty measures related to the degree of disclosure for any unit in the population and how to specify a lower uncertainty limit. Heldal and Fosen (2001) are also referring to literature that the critical kind of disclosure to avoid is identity disclosure, which is often a prerequisite for attribute disclosure.

Another issue related to statistical inference is cases with sporadic observations. Rajasekaran et al. (2009) put forward the idea that observations far from the mean have a high risk of being disclosed and thereby introduce a rationale for suppressing cells with few observations. This can, for example, be the case for commuting between zones involving a long distance.

2.2 Balancing Confidentiality and the Utility of Research

As stated above, a tradeoff between privacy protection and accuracy represents an optimization problem. Abowd and Schmutte (2019) aimed at developing a principled framework for dealing with this tradeoff. One component of this framework is a "closed, bounded, and convex production function relating privacy loss and statistical accuracy", (Abowd and Schmutte, 2019). This production frontier defines a marginal rate of transformation. At the same time, the specification of preferences introduces the marginal rate of substitution between privacy loss and accuracy, also referred to as "the willingness to accept privacy loss measured in units of statistical accuracy", (Abowd and Schmutte, 2019). In a normative approach, the statistical agency then maximizes a utilitarian social welfare function, given the restriction represented by the production frontier. This leads to an optimal combination of privacy loss and accuracy, corresponding to the condition that the marginal transformation rate equals the marginal willingness to accept privacy loss. According to Abowd and Schmutte, 2019), this optimization approach is "generally valid for all differentially private mechanisms that yield a convex relationship between privacy loss and accuracy", and they prove that an approach with randomized response gives a strictly increasing and concave relationship between privacy loss and accuracy.

Abowd and Schmutte, 2019 further argue that the accuracy of published statistics can be seen as nonrival and non-excludable in consumption and hence considered a public good. Similarly, it is argued that the differential privacy parameter ε reflects a social issue and that an SDL procedure gives the entire population the same protection against privacy loss. This means that privacy protection is also strictly nonrival and can be considered a public good; for instance, the discussion provided by Jones and Tonetti (2020). Since statistical accuracy and privacy are public goods, their optimal levels are a social choice (Abowd and Schmutte, 2019). As mentioned, Abowd and Schmutte (2019) provide a normative approach, and they claim that making their framework practical calls for more sophisticated models of production possibilities and better models and measures concerning the demand for privacy and accuracy (Abowd and Schmutte, 2019). They also claim that statistical agencies should aim at solutions where the marginal costs are equal to the marginal benefits of statistical disclosure limitation procedures. However, "statistical agencies are not yet using formal privacy protection systems", (Abowd and Schmutte, 2019), and not much research is provided on such issues.

Abowd and Schmutte (2016) point out that the effects of SDL on the results are more severe when the analysis aims at explaining the situation for a specific subpopulation that is in particular exposed to the confidentiality issue. They further define SDL to be ignorable if "the analysis can recover the estimates of interest and make correct inferences using the published data without explicitly accounting for SDL" (Abowd and Schmutte, 2016). In general, SDL may cause a

severe bias in a rural-urban dimension, reflecting the likely event (Carpenter et al. (2022)) that less-populous geographies have a higher share of suppressed cells. Another exciting perspective concerning the tradeoff between privacy and the usefulness of data is related to the property rights for data. Jones and Tonetti (2020) addresses the possibility that consumers' privacy is not adequately accounted for if firms own the data. At the same time, the nonrival character means potentially substantial social gains in making data broadly accessible.

2.3 Alternative Ways of Limiting Disclosure

One obvious step to avoid disclosure of private information is removing data that directly allows identification, like name and home address. However, this is generally insufficient to maintain the individual's privacy, as demonstrated in Sweeney (2002). Hackers may, for instance, combine different databases to reach sensitive information, and in general, a small number of demographic attributes allows the identification of an individual (Abowd and Schmutte, 2016). If a specific combination of residence and job involves very few workers, there is a high probability that the workers can be identified. This also applies if some individuals in a database have some easily recognizable, extreme values of some variables. A database of location information, such as an origin-destination matrix of commuting, can serve as an example of data that can be used to disclose more sensitive individual information from other databases.

Matthews and Harel (2011) provide a review of methods employed for maintaining the privacy of publicly released data, and a comprehensive review can be found in G. T. Duncan et al. (2011). Entering into details on different methods is beyond the scope of this paper. However, the following list of methods, based on Matthews and Harel (2011), may serve as a valuable backdrop for the discussion to follow:

I. Basic methods

- Limitation of detail; variables can be recorded into intervals, and/or categories can be collapsed together.
- Top/bottom coding; a variable's largest or smallest value is limited.
- Suppression; cells with too few observations in a contingency table are not released to the public but replaced with missing values. This is "one of the most common forms of SDL" (Abowd and Schmutte, 2016), and "... all such public data have been subjected to very substantial SDL, almost all of it in the form of suppression" (Abowd and Schmutte, 2016).
- Rounding; each observation is rounded up or down to the nearest multiple of the rounding base; rounding up or down is decided upon randomly.

- Addition of noise; noise can be added to the data to prevent identification through linkages. Noise infusion inflates variances. If several releases of information are based on the same method, the noise distribution can be inferred (Abowd and Schmutte, 2016).

II. Sampling; if only a microdata sample is released, potential intruders cannot be sure that a unique match from another dataset identifies an individual.

III. Matrix masking; appropriate conformable matrices are released. This includes noise addition, sampling, suppressing sensitive variables or cells, adding simulated data, and all exceptional cases of matrix masking. Reliable analyses call for knowledge of the masking procedure. Another option is to use randomized responses to gain sensitive information. The sensitive question can be given with a probability p, while there is a probability of $1 - p$ that the respondent is given a trivial question. The sensitive information then remains private even in cases of identification of individuals.

IV. Data swapping; as stated by Dalenius and Reiss (1982), the original data matrix is used as an "input" for producing another data matrix, "which is used as the basis for producing statistics by way of tabulations". According to Abowd and Schmutte (2016), "Data swapping is the practice of switching the values of a selected set of attributes for one data record with the values reported in another record". Individual information is masked, the marginal counts of the contingency table are maintained, and the data-swapping only needs to be done on sensitive variables by changing units between cells.

V. Synthetic data; observed microdata can be considered as a random sample from a population. The unsurveyed part of the population is assigned plausible values of sensitive data through a technique with multiple imputations generating randomly drawn data from a posterior predictive distribution, which means that they are based on information from the sampled part of the population. Eventually, only data for the unsurveyed part of the population may be released. This idea of generating synthetic data for statistical disclosure limitations was introduced by Rubin (1993). Abowd and Schmutte (2016) introduces the possibility that synthetic data are validated by the data providers on the actual confidential data. As explained in Section 3, we generate synthetic data to discuss challenges and corrections for suppressed data in a commuting context.

VI. Aggregation and other selected methods; slicing complete data into groups involving a smaller number of variables, while microaggregation is based on creating new records by averaging at least three original records. Geographic units can be aggregated, data on occupation and industry can be released in broad categories, and incomes can be reported in bins (Abowd and Schmutte, 2016). Matthews and Harel (2011) in addition, discuss

methods to limit disclosures related to location data collected from commuters' global positioning system (GPS). They also discuss Argus, a software package for limiting the potential for statistical disclosure by limiting the occurrence of rare combinations of variables.

VII. Micro-agglomeration, substitution, subsampling, and calibration; is a combination of different statistical disclosure techniques, proceeding in four steps, followed by calibrating the released data such that specific estimates based on released data match the estimates from the observed data.

This list of methods to deal with SDL refers to the use of microdata. If such methods are applied to microdata, they also represent a source of error in tabular data. This is particularly unfortunate if the analysis focuses on issues sensitive to observations in cells representing choices made by a few individuals in the population. According to Abowd and Schmutte (2016), the most common method to deal with sensitive cells in a table is suppression, while randomized rounding is also used in many cases. They further claim that adding noise to the microdata may be a preferred alternative to suppression. Noise infusion means that variance will be added to published data while reducing bias. Chetty and Friedman (2019) discuss how the risk of privacy loss can be reduced by adding noise to parameter estimates.

Abowd and Schmutte (2016) also explain the need for complementary suppression for tabular data. This is, for instance, due to the possibility that users can deduce values of sensitive cells from information on, for example, marginal sums in the table. In other words, a residual disclosure may find a place as stated in Heldal and Fosen (2001). Statistics Norway demands that at least three units have to underly a total value to be published in a cell to avoid the risk of a residual disclosure. In many countries, this is not considered sufficient, restricting that the three main contributors to the cell total represent at least 80% of the total sum. This is a case of the so-called (n,k) dominance rule, suppressing the cell if the n largest units contribute more than k % of the total sum. This dominance rule is primarily introduced to keep vital information hidden for competing firms, for instance, concerning variables like production, sales, wage costs, import, export, etc. Still, this information may be available from other sources, and official statistics will probably not be the first source of information where competing firms search for relevant information.

Another approach to avoid that information in the suppressed cells can be deducted from the information of the marginal totals in a table is to suppress the information in more cells; that is a secondary suppression Heldal and Fosen (2001). Heldal and Fosen (2001) discuss τ-ARGUS as a program constructed to optimize the cell suppression in cases where the marginal total in a table is known through a secondary suppression to avoid a residual disclosure. Microdata, on the other hand, have no marginals, and sensitive values should be substituted by "missing" (Heldal

and Fosen, 2001). An aggregated table with complete cell information can then not be correctly reproduced from microdata.

One challenge for researchers is that data publishers do not, in general, disclose the methods used for complementary suppression. However, Abowd and Schmutte (2016) claim that agencies providing data are becoming more open to using noise-infused methods in producing data tables. Adding noise to statistics generated from a database belongs to a general class of approaches called matrix mechanisms, which are discussed in detail in Li et al. (2015). According to Carpenter et al. (2022), US statistical agencies use cell suppression and noise infusion, "with some published cell values perturbed by a random noise multiplier, to prevent the disclosure of individual business establishment information" (Carpenter et al., 2022).

An approach with noise infusion may also be performed by adding noise to the estimates from the regression based on a small sample. This technique is based on the differential privacy concept discussed in Subsection 2.1. Chetty and Friedman (2019) compare such a procedure to a standard cell suppression approach. Their comparison can be summarized in terms of the following three dimensions:

Privacy loss The noise infusion approach introduced by Chetty and Friedman (2019) substantially reduces the risk of privacy loss compared to cell suppression. Chetty and Friedman (2019) claim this will be the case for most noise infusion approaches.

Statistical bias The parameters underlying the random noise infusion are publicly known (Chetty and Friedman (2019)). Hence, unbiased parameters can be reached, as opposed to what, in general, is following from the measurement errors resulting from approaches based on count-based suppression.

Statistical precision Estimates based on noise infusion are less precise than those following cell suppression methods. According to Chetty and Friedman (2019), this is the key drawback of noise infusion and the primary concern of most researchers.

As pointed out by Abowd and Schmutte (2016), suppression represents a missing data problem that can be treated with ad hoc methods. One is to ignore the suppressions, and another is to analyze a more aggregate level. A better solution (Abowd and Schmutte, 2016) is to allocate values of high-level aggregates into missing cells at a lower level of aggregation. As the best approach, Abowd and Schmutte (2016) proposes to combine the relevant model with a model for suppressed data, for instance, by a Bayesian hierarchical model.

Before the data release, many different approaches had been proposed for statistical disclosure limitations (SDL). In many cases, disclosure limitations will not be expected to affect the results of empirical analysis. According to Abowd and Schmutte (2016), modifications due

to confidentiality concerns are generally modest relative to other more serious data quality problems, like reporting errors and missing items.

2.4 Accounting for Statistical Disclosure Limitation Methods

Suppose methods for statistical disclosure limitations prove to be non-ignorable in terms of affecting the results. In that case, they should be explicitly adjusted for the analysis to avoid biased estimates of parameter values and/or the corresponding variances. Such adjustments call for knowledge of the SDL methods used by the data publisher or methods that can recover the SDL parameters from prior information and the released data. From a researcher's point of view, it would be valuable if details on the appropriate SDL method were made public by the data-releasing agencies.

Abowd and Schmutte (2016) provide two examples of such methods. One is an approach with randomized responses to sensitive issues; the other is top coding of incomes, censoring incomes above a specific threshold. In both cases, the SDL method obviously should be accounted for in analyzing the data; see Abowd and Schmutte (2016) for a discussion. Abowd and Schmutte (2016) also discuss how SDL methods may lead to attenuated parameter estimates and underestimated standard errors in linear regression models. They refer to analyses addressing the closely related biases resulting from other sources of missing data, like missing responses to specific questions in a survey. Abowd and Schmutte (2016) provide a discussion of how the bias resulting from SDL can be corrected. One approach is to use the information on the noise variance in a case where this is the relevant SDL and then correct the bias analytically. As another approach, it is well known that finding an appropriate instrument may be an appealing solution to deal with measurement errors (Abowd and Schmutte, 2016, Carpenter et al., 2022). The challenge is finding an instrument correlated with the true but unknown data but not with the suppressed information representing the measurement error.

Abowd and Schmutte (2019) distinguish between a reconstruction attack and a re-identification attack. The former is about building a copy of a confidential database based on statistics produced and published from this database. If many linearly dependent statistics are available, there may be a substantial potential for reconstructing hidden variables, which is a data breach. Marginal totals from a contingency table are, for example, linear statistics that can be used for reconstruction. Abowd and Schmutte (2019) provide an example from county tables on the Quarterly Census of Employment and Wages, published by the Bureau of Labor Statistics, concerning the information of payroll and employment data of business establishments. Based on information from summaries, the suppressed cells can be reconstructed very accurately from information in the time series of the county tables (Abowd and Schmutte, 2019). Re-identification involves linkage to information from other external databases, deterministically or

probabilistically, see for instance Airoldi et al. (2011).

Carpenter et al. (2022) discuss an increasing demand for private data sources. Private agents may provide data where suppressed cells are estimated, for example, through the availability of new data, like online business directories, underpinned by increased computational power. According to Carpenter et al. (2022) and Abowd and Schmutte (2019) this may lead to privacy concerns and to a higher percentage of suppressed cells. Hence, an increasing interest can be expected in how cell suppression can invalidate statistical inference and how this should be treated analytically. Carpenter et al. (2022), claim that most journal articles in economic literature do not document the methods used in estimation based on suppressed cell data sets. One possible technique, mentioned by Carpenter et al. (2022), is an iterative proportional fitting procedure, which is "an algorithm for estimating cell values of a contingency table such that the totals remain fixed and the estimated table decomposes into an outer product" (Carpenter et al., 2022, p 59). This technique has been used (Bartik et al., 2018), resulting in estimates labeled WholeData, which are free to researchers. Carpenter et al. (2022) use WholeData to examine measurement errors resulting from suppressed cell data sets, and they claim that many firms offering purchasable data use nonpublic algorithms to provide suppressed cell data estimates. Guldmann (2013) provides a review of methods to deal with suppression issues. According to Carpenter et al. (2022) "methods doubtless vary in their success, but none address the larger question of how much error remains and how that error influences economic and statistical significance". It is a challenge that missing information on such methods limits both validity testing and research reproducibility (Carpenter et al., 2022).

3 Generating a Synthetic Population and Data for the Analysis

As was clear from Section 2.3, one possible statistical disclosure limitation method is generating a synthetic population. Commensurate with the fundamental problem being discussed in this paper, we do the analysis based on a synthetic population. Hence, we are not faced with confidentiality issues, apart from reaching insight on spatial interaction behavior that may contribute to identifying population subgroups.

A synthetic population can be generated through agent-based modeling. As pointed out by A. Wilson (2010), this is represented by a system where the individual agents make decisions according to probabilistic rules of behavior. The observed spatial pattern results from decentralized decision-making, as demonstrated in for example Page (1999), Anas (1983a) and Irwin (2010). The construction of such a pattern is essentially reflecting two separate steps. The first specifies the agents' geography and population, while the second introduces the labor and housing market

conditions that define the opportunity set of the utility-maximizing agents. For an example of how such a modeling framework is parameterized and applied, see Gholami et al. (2022).

3.1 Specifying the Geography and Generating a Population

The geography is represented by the 12-node system illustrated in Figure 1. Notice from the figure that there are three clusters of towns. The towns A, B, C, and D are located on another side of a topographical barrier than the other towns. There are short distances between the towns within each cluster but relatively long distances across the different clusters. This means that a relatively high number of cells in the matrix can be expected to involve a low number of commuters, leaving an imminent risk of identification disclosures.

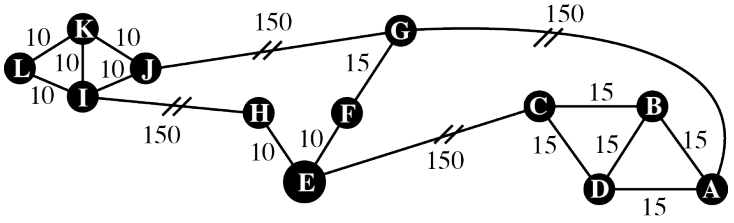

Figure 1: The geography of towns. The numbers marked in the figure represent the kilometers' distances for the corresponding segments in the road transportation network.

For technical convenience, similar to the construction in McArthur et al. (2010), McArthur et al. (2012), the generation of a population is initiated with a large number of fifteen-year-old agents, who are distributed between the 12 towns and by gender according to given probabilities. The population members then interact according to rules based on Norwegian statistical data. This involves probabilities of getting married, having children, being divorced, being retired, dying at a specific age, etc. Some rules have a spatial dimension beyond the general probabilities defined by observed frequencies in the Norwegian population. This applies, for instance, to the probability of getting married. Spouses are drawn from the population of single people, but a distance-related deterrence effect is introduced in the linking to a partner. The probability of getting married to a person is assumed to decline exponentially with the distance to his or her residence, which means that a parameter is introduced reflecting the tendency to end up with a local partner. More details of the demographic rules underlying the formation of a population can be found in McArthur et al. (2010), McArthur et al. (2012), who consider a simple two-node geography.

The interaction succeeding the initialization goes on for 300 years. After this period, there are no more traces of the initial population, and we are left with a population that mimics the Norwegian in a demographic sense. Our microsimulation experiments start in year 300,

resulting, for instance, in a pattern of commuting flows corresponding to individual decisions determined by preferences and budget considerations.

Our agents are equipped with preferences. The preferences for each agent are specified as a Cobb Douglas function of their housing consumption and the consumption of other goods and services. This involves a parameter reflecting the relative weight put on these aggregates in the preferences. For each person, the value of this parameter α_i is drawn randomly from a normal distribution with $\alpha_i \in (0.3, 0.7)$.

As a next step, we are specifying the economic system, including the opportunity set of the utility-maximizing agents. The total consumption and the utility level depend on an individual's lifetime earnings, which mirror the labor and housing market situation in the different towns of the region.

Workers apply for jobs that contribute to increasing the utility level of the household. There are three categories of workers and three categories of related jobs. Within each group, the workers are assumed to be homogeneous regarding job qualifications. Applicants are randomly selected for vacant positions, and a job offer means they can stay in this position until they retire.

If a worker accepts a job offer from a firm located in another town than his/her current residence, there are two possible forms of spatial interaction response. One is to move to the town hosting the new employer, and the other is commuting. The chosen alternative is the one that maximizes the sum of utilities of the spouses.

There are two kinds of firms. In line with economic base modeling, employment in basic sector firms is assumed to be exogenously given, while local sector firms serve the population in the region. There is a cluster of local sector firms in Town E, which is the central business district, attracting customers from other towns in the region (Gjestland et al., 2006). This generates commuting flows between the towns, as does the friction represented by the heterogeneity of job and worker categories. Both this tendency to shop in the regional center and the exogenous spatially uneven distribution of basic sector jobs represent an unbalance in the residential and job location pattern.

Lifetime earnings reflect wages and insurance payments when an individual is unemployed. Due to a hypothesis of agglomeration economies, wages are initially assumed to be higher in a regional center, town E, than in the other towns of the region. For the years to follow, spatial wage disparities may follow from local Phillips style mechanisms, see McArthur et al. (2010), McArthur et al. (2012).

McArthur et al. (2010), McArthur et al. (2012) did not account explicitly for how spatial disparities in housing prices may result from the balance of supply and demand in the local

housing market. Spatial disparities in housing prices may strongly influence the decision to move or commute as a response to a shift in the job location. The housing demand is represented by first-price sealed bid auctions, where all bids are submitted simultaneously, and the highest bid wins the auction. Our model has a pool of houses for sale, and agents make a random check into this pool to see if they may join the pool of potential bidders. All agents in the pool scan the pool of houses for sale, searching for the one that provides them with the highest utility. A sale is completed if a seller finds the winning bid exceeds his or her reservation price. If the sale is not completed, the house remains in the pool of houses for sale. For more technical details on this bidding procedure, see Gholami et al. (2022).

High prices can be expected locally if there is a high demand for houses in a particular year. The initial housing supply is assumed to be provided by a regional government planning entity, and the building frequency of new houses is assumed to be a minimum of 5% yearly in each town. In our model, high housing prices in the previous year serve as an incentive and explanation for building frequencies exceeding 5% in the current year.

3.2 Adding Workers to the System

As stated in Section3.1, the geography is constructed to reach a solution where a relatively high number of cells in the matrix can involve a low number of commuters. In our microsimulations, this can be reinforced by the specification of wages, housing prices, and other parameters that influence individual spatial interaction decisions. Hence, the case we consider is constructed to reach a state where a high number of town combinations involve a low number of commuters, inducing confidentiality issues.

Through the generation of the synthetic population, we know the state of each agent at any time, represented by their family relations, preferences, residential location, and work location. This means that the corresponding commuting flow pattern results from utility-maximizing agents' decisions. To get a higher diversity of low numbers in the cells of the trip distribution matrix, we systematically add some long-distance commuters to the system. This further contributes to introducing exciting information regarding the time-series properties of the commuting pattern. As we will demonstrate in Section 5, time-series data proves to help disclose information suppressed by the data publisher. For example, one kind of relevant information is represented by cells where the value changes over time. This may cause transitions from years where the cell information is suppressed to years where this particular cell has a value that is released.

Our relatively small system, with only 12 towns, has relatively few cells where the value changes considerably over time. Adding a few long-distance commuters extends the potential to study the relevance of time-series effects. We can, for instance, add 1 or 2 workers to all the cells where 0 or 1 workers are observed in the commuting trip matrix following directly from our

microsimulations for year 300. These workers are assumed to keep their initial combination of residence and job location for the entire period under study. This defines four scenarios for year 300 and four concerning the number and time-series profile of suppressed cells, with the values changing more frequently between 0, 1, 2, 3, 4, etc. As reported in Section 5, we do experiments with all these scenarios.

These added workers are, of course, not rational agents making utility-maximizing decisions, and they make up a deficient proportion of the otherwise rational synthetic population. This approach resembles the introduction of so-called noise traders in finance, that is, traders making irrational decisions, trading by random, deviating from market averages (see, for instance, https://www.investopedia.com/terms/n/noisetrader.asp). Adding a few irrational, uninformed workers improves the potential of studying the effect of dynamics and transitions in adjusting for statistical disclosure limitations. It also opens for more detailed studies of finding an appropriate cut-off value in a strategy where cell suppression is the chosen SDL approach.

3.3 Non-Suppressed and Suppressed Trip-Distributions

For each year, there is a total of l_i workers in town i, e_i working places, and we let T_{ij} denote the number of workers who live in town i and work in town j. The actual trip-distribution matrix is a 12×12 matrix. An example of an actual trip-distribution is shown in Table 1, which refers to year 17 of a 20-year long time period following from the initialization of our agent-based generation of the population. Our approach works particularly well for year 17, which makes this year appropriate in clarifying what we aim to achieve by the procedure described in Section 4.2. The sample results from the procedure described in Sections 3.1 and 3.2. In the following we will refer to the actual trip-distribution as the non-suppressed matrix.

Table 1: Non-suppressed trip-distribution for year 17.

	A	B	C	D	E	F	G	H	I	J	K	L
A	534	96	65	98	89	3	2	2	1	1	1	1
B	107	372	112	100	91	1	3	0	1	1	1	1
C	93	92	390	104	114	5	1	4	1	3	1	0
D	108	93	142	339	104	1	2	1	1	0	1	2
E	5	3	3	5	1470	190	144	288	3	1	3	1
F	2	2	4	2	623	228	103	42	2	2	3	1
G	2	2	1	1	433	26	206	20	1	1	2	1
H	2	3	2	4	555	45	61	231	2	1	1	1
I	1	1	1	1	114	1	3	2	294	130	111	121
J	2	1	2	1	109	4	2	2	138	464	145	69
K	1	1	1	1	83	3	1	4	107	102	348	102
L	1	1	1	1	114	2	1	8	141	96	122	293

The matrix of commuting trips in Table 1 reflects the spatial configuration shown in Figure 1, with a long distance between the three groups of towns. There are many cells with a low number of commuters, but these workers do not make up a high proportion of the total number of commuters. The commuting from towns G, H, I, J, K, and L into town A will be suppressed if a cut-off value of 3 is used. This involves 11 commuters, which make up approximately 1.28% of the total number of jobs in town A and approximately 3.4% of the total in-commuting towards town A. The figures are roughly the same for the other towns, except for the regional center E, for which no information about in-commuting is suppressed. The suppressed trip-distribution is shown in Table 2.

Table 2: Suppressed trip-distribution for year 17.

	A	B	C	D	E	F	G	H	I	J	K	L
A	534	96	65	98	89	3	0	0	0	0	0	0
B	107	372	112	100	91	0	3	0	0	0	0	0
C	93	92	390	104	114	5	0	4	0	3	0	0
D	108	93	142	339	104	0	0	0	0	0	0	0
E	5	3	3	5	1470	190	144	288	3	0	3	0
F	0	0	4	0	623	228	103	42	0	0	3	0
G	0	0	0	0	433	26	206	20	0	0	0	0
H	0	3	0	4	555	45	61	231	0	0	0	0
I	0	0	0	0	114	0	3	0	294	130	111	121
J	0	0	0	0	109	4	0	0	138	464	145	69
K	0	0	0	0	83	3	0	4	107	102	348	102
L	0	0	0	0	114	0	0	8	141	96	122	293

A zero entry in a suppressed matrix of commuting trips, e.g., Table 2, means that the actual value can be 0,1, or 2, assuming a cut-off value of 3. The basic approach taken in the paper is that the total number of workers and working places in each town are public, and our challenge is hence to offer a prediction of the actual trip distribution based on the suppressed data in Table 2 and the known marginal totals from the non-suppressed matrix of commuting trips in Table 1.

Table 3 provides the results of our disclosure procedure in a case with a cut-off value of 3. To demonstrate how discrepancies cancel each other out to reach a balanced trip matrix, the example in Table 3 corresponds to a case where our optimization procedure provided a close to perfect reproduction of the true trip distribution. This stems from using the objective function presented in Section 5.2, and the results for year 17 in Table 4.

Table 3: Inferred trip-distribution for year 17 using the objective function presented in Section 5.2.

	A	B	C	D	E	F	G	H	I	J	K	L
A	534	96	65	98	89	3	2	2	1	1	1	1
B	107	372	112	100	91	1	3	0	1	1	1	1
C	93	92	390	104	114	5	1	4	1	3	1	0
D	108	93	142	339	104	1	2	1	1	1*	1	1*
E	5	3	3	5	1470	190	144	288	3	1	3	1
F	2	2	4	2	623	228	103	42	2	1*	3	2*
G	2	2	1	1	433	26	206	20	1	1	2	1
H	2	3	2	4	555	45	61	231	2	1	1	1
I	1	1	1	1	114	1	3	2	294	130	111	121
J	2	1	2	1	109	4	2	2	138	464	145	69
K	1	1	1	1	83	3	1	4	107	102	348	102
L	1	1	1	1	114	2	1	8	141	96	122	293

By comparing Table 1 with Table 3, we see that we have a near-perfect match, the only difference is the four entries marked out with $*$. It is straightforward to see that the discrepancies cancel each other out regarding the marginal sums of rows and columns. The prediction of 1 rather than 0 commuters from town D to town J is balanced by the prediction of 1 rather than two commuters from town D to town L. Hence, the row sum will be left unchanged. Column-wise, the prediction of 1 commuter from town D into town J shows that the predicted commuting from town F to town D is one rather than the true value of 2 commuters. A similar accounting applies to column L.

4 Entropy

Throughout this paper, we will base our predictions on special cases of entropy-maximization. Several authors have studied entropy methods in connection with suppressed data, but to the best of our knowledge, this paper is the first to discuss such methods in the context of trip distribution.

A. Wilson (2010) claims that entropy-maximizing models are appropriate for estimating missing data, for instance, when row- and column sums are known. Airoldi et al. (2011) propose an entropy metric for assessing how the risk of trail disclosure depends on the distribution of how people visit different sets of locations that may share their information for linkage purposes. Through case-based and controlled experiments, Airoldi et al. (2011) demonstrate that the entropy metric effectively estimates the risk of trail disclosures, suggesting that low entropy systems correlate with high re-identifiability. In general, entropy-based measures have been used

to estimate information loss (Willenborg and De Waal, 2012) and also for estimating disclosure risk (Bezzi, 2007). Rodrigues (2016) uses the maximum entropy principle to determine stochastic properties of data where uncertainty results from suppression or other kinds of measurement errors. In the specification of the optimization problem, Rodrigues (2016) accounts for the possibility that there are correlations in the probability distribution of missing observations. Such correlations follow accounting identities, like the marginal totals of a trip distribution problem. Majeed and Lee (2020) focus on the users' privacy in social networks. They estimate the level of uncertainty for disclosure by computing the entropy of sensitive attribute values, with high values of entropy meaning that there is a substantial potential for protecting the privacy of the users of social networks. Based on maximizing the entropy of sensitive attribute values, Tsiafoulis et al. (2011) are building equivalent classes with uniformly distributed values of such attributes before noise is introduced to preserve privacy. The anonymity is better, and identity disclosures are more difficult for high entropy values. Building equivalent classes with uniform distributions of sensitive attributes minimizes the required noise and information loss. González-Vidal et al. (2020) demonstrate that a framework with Bayesian maximizing entropy is generally very well suited to deal with missing value imputation in Internet of Things applications.

Entropy considerations date back to the classical works of Boltzmann. Boltzmann's idea was to count how many ways a given macro-state can be realized in terms of randomly selected micro-states, and he assumed that the most likely macro-state was the one that could be realized in the largest number of ways. Shannon entropy, also called information uncertainty, works similarly. In both cases, the entropy H is defined by

$$H = -\sum_{i=1}^{N} P(x_i)\log(P(x_i))$$

where $x_1, x_2, ..., x_N$ are the possible microstates and for $i = 1, \ldots, N$: $P(x_i)$ is the probabilities that the microstate x_i is selected. In the context of trip distributions, the agents select where to live and where to work. In a system with n towns, we get $N = n^2$ possible microstates. If we let π_{ij} denote the probability that an agent selects to live in town i and work in town i, the entropy is

$$H = -\sum_{i,j \in I} \pi_{ij}\log(\pi_{ij}).$$

where $I = \{1, 2, \ldots, n\}$ is the index set of the n towns. If M is the total number of agents living in the system, we expect $T_{ij} = M\pi_{ij}$, and without loss of generality, we can define $\pi_{ij} = \frac{T_{ij}}{M}$. If we rewrite the entropy noting that (by definition) $\sum_{i,j \in I} T_{ij} = M$, we get

$$H = -\sum_{i,j\in I} \frac{T_{ij}}{M} \log\left(\frac{T_{ij}}{M}\right) = -\frac{1}{M}\sum_{i,j\in I} T_{ij}\log(T_{ij}) + \frac{1}{M}\sum_{i,j\in I} T_{ij}\log(M)$$
$$= -\frac{1}{M}\sum_{i,j\in I} T_{ij}\log(T_{ij}) + \log(M)$$

We hence see that in a system with a fixed number, M, of agents, maximum entropy is obtained when $-\sum_{i,j\in I} T_{ij}\log(T_{ij})$ is as large as possible.

4.1 Solution to Constrained Maximum Entropy Problems

To infer the most likely trip distribution under marginal constraints, we search for the solution to the following optimization problem:

$$\max_{T_{ij}} -\sum_{i,j\in I} T_{ij}\log(T_{ij})$$

under the constraints

$$\sum_{i\in I} T_{ij} = e_j \qquad \sum_{j\in I} T_{ij} = l_i$$

This problem is easy enough to admit a closed-form solution, and it is straightforward to see that maximum entropy is obtained when

$$T_{ij} = A_i B_j$$

where $A_1,\ldots,A_n,B_1,\ldots,B_n$ are constants. Numerical values for these constants can be found from the balancing constraints

$$\sum_{i\in I} A_i B_j = e_j \qquad \sum_{j\in I} A_i B_j = l_i.$$

Efficient numerical algorithms, e.g., the Bregman algorithm, Bregman (1967), can be used to find these constants, see Sen and Smith (1995). The solution to the maximum entropy problem coincides with the expectations in a Chi-square table under independence.

The problem becomes more interesting if we impose further constraints. In the trip-distribution problem, it is natural to assume a generalized cost c_{ij} when an agent commutes between the origin i and the destination j. Such generalized costs can include start-up costs, driving distance, travel time, etc. Assuming that the total generalized commuting cost C is given, we might consider a modified maximum entropy problem, i.e.

$$\max_{T_{ij}} -\sum_{i,j\in I} T_{ij}\log(T_{ij})$$

under the constraints

$$\sum_{i\in I} T_{ij} = e_j \qquad \sum_{j\in I} T_{ij} = l_i \qquad \sum_{i,j\in I} T_{ij}c_{ij} = C$$

This modified problem, too, admits a closed form solution:

There exists constants $A_1,\ldots,A_n,B_1,\ldots,B_n$, and a unique constant β such that

$$T_{ij} = A_i B_j e^{-\beta c_{ij}} \tag{1}$$

The solution to the modified maximum entropy problem is a multinomial logit model. It is interesting to note that the solution coincides with the solution to the random utility problem, i.e., the resulting distribution we get when agents select origins and destinations subject to random utility maximization, McFadden (1974) and Train (2003).

The model in (1), can be derived in several additional ways, it can be derived from maximum entropy considerations, see, e.g., A. G. Wilson (1967), Anas (1983b), Erlander and Stewart (1990). It is the solution of the maximum utility problem, Erlander and Stewart (1990), and a consequence of probabilistic cost efficiency, Erlander (2010). In addition there exist several other approaches, e.g., Mattson and Weibull (2002) and Matějka and McKay (2015), leading to the same model. Contrary to the many economics models chosen for analytical convenience, the approach leading to (1) is hence firmly anchored in statistical theory.

4.2 Inferring Suppressed Data Using Maximum Entropy Methods

In the previous section, we have seen that a maximum entropy approach produces models consistent with models widely used in the literature on trip distribution. To proceed to cases with suppressed data, we let $\mathscr{S}$ denote the index set of the suppressed entries. We then consider a modified maximum entropy problem:

$$\max_{T_{ij}} -\sum_{i,j\in I} T_{ij}\log(T_{ij})$$

under the constraints

$$\sum_{i\in I} T_{ij} = e_j \qquad \sum_{j\in I} T_{ij} = l_i$$

and where

$$T_{ij} = \begin{cases} T_{ij}^{\text{non-suppressed}} & (i,j) \notin \mathscr{S} \\ \text{a free optimization variable} & (i,j) \in \mathscr{S} \end{cases}$$

We fix all the unsuppressed entries and use entropy maximization to infer the suppressed values. In problems of this kind we use the convention that $x\log(x) = 0$ when $x = 0$, which comes as a natural consequence of the continuous limit $\lim_{x\to 0^+} x\log(x) = 0$. Computer software like AMPL can easily solve integer-constrained optimization problems of this sort.

In general, we will report the fit in terms of accuracy, which we define as follows:

$$\text{accuracy} = \frac{\text{\# correct values}}{\text{\# suppressed values}}$$

In our analysis, we will also report the fit in terms of SRMSE over the suppressed entries, which we define as follows:

$$\text{SRMSE} = \frac{\sqrt{\dfrac{\sum_{(i,j)\in\mathscr{S}} (T_{ij} - T_{ij}^{\text{non-suppressed}})^2}{|\mathscr{S}|}}}{\dfrac{\sum_{(i,j)\in\mathscr{S}} T_{ij}^{\text{non-suppressed}}}{|\mathscr{S}|}}$$

i.e., we compare the mean square deviation over the suppressed entries with the mean of the actual values that have been suppressed. Here $\mathscr{S}$ are the suppressed entries, and T_{ij} is the solution to our modified maximum entropy problem. The SRMSE is a dimension-free quantity that allows us to compare cases where the number of suppressed values is very different.

The example considered in Table 2 was used for demonstration purposes. To examine the performance in more general cases, based on a synthetic population for the 12-node geography illustrated in Figure 1, we started with a pure entropy maximization approach without accounting for additional information on the system. This is done for 20 years, and we generated a time series of non-suppressed matrices from year 1 to year 20.

We believe our procedure is neutral in that there is no part of the construction we expect would favor maximum entropy methods. In the time-series approach discussed in the next section, however, we wanted a data set where several cells flip between suppressed and public over time. As the original data set contained few such cases, we traced the cells with 0 or 1 commuters in year one and added one non-optimizing worker in each cell. This worker kept the initial

combination of residence and job location for the entire period. As explained in Section 3.2, we experimented with three other ways of adding workers but settled with the procedure detailed above.

Entropy maximization, in general, works fine in disclosing suppressed information on commuting trips. As an average over the 20-year-long period, the accuracy is found to be 0.7537, which means that around 3 out of 4 suppressed cells are perfectly disclosed by maximizing the entropy. In comparing the number of disclosed commuters in the suppressed cells to the known actual numbers, an average value of 0.3913 was reached for the SRMSE. For the performance of this pure entropy maximization approach for specific years over the period, see the first five columns of Table 4.

5 Extension to Time Series

The results presented in the previous section demonstrated that entropy maximization might perform exceptionally well in disclosing information suppressed in the matrix of commuting trips. One interesting question is whether information on the suppressed trip distribution at several consecutive points in time can be utilized to give even more accurate estimates of the suppressed information. Does time series data comprise a potential for further improvements in disclosing suppressed data?

Our core model extends trivially to the time series case. We let $\mathscr{T}$ denote the index set of times, and for each $t \in \mathscr{T}$ we let $\mathscr{S}_t$ denote the index set of the entries in the trip-distribution that are suppressed at time t. We then consider

$$\max_{T_{ijt}} - \sum_{i,j \in I, t \in \mathscr{T}} T_{ijt} \log(T_{ijt})$$

under the constraints

$$\sum_{i \in I} T_{ijt} = e_{jt} \qquad \sum_{j \in I} T_{ijt} = l_{it}$$

and where

$$T_{ijt} = \begin{cases} T_{ijt}^{\text{non-suppressed}} & (i,j) \notin \mathscr{S}_t \\ \text{a free optimization variable} & (i,j) \in \mathscr{S}_t \end{cases}$$

This extension is straightforward but does not add anything in terms of improved performance. The reason is that the different years are effectively disconnected in this model, and maximum entropy is obtained by finding maximum entropy for each year separately.

To improve performance, we need to modify the objective function to consider the time develop-

ment. A central idea is that when a value flips from zero to a public value or from a public value to zero, it is likely that the suppressed value is close to the cut-off. Such cases only involve a small number of agents, and the effect is typically triggered when one agent randomly changes status. It seems less likely that several agents change their status simultaneously. Time series information of this sort can be implemented in many different ways, and in the next section, we discuss some ways of doing this.

5.1 Incorporating Time Series Information

We first choose and fix the entry (i, j) in the trip distribution matrices, and consider what happens to the flow T_{ijt} when $t \in \mathscr{T}$. For some values of t, it may happen that T_{ijt} is suppressed, while at other times, this flow might not be suppressed. We define a new reward parameter f_{ij} as follows:

$$f_{ij} = \frac{\text{\# of years where } T_{ijt} \text{ is not suppressed}}{\text{\# years in total}}$$

The intuition behind this parameter is as follows: the more years we have unsuppressed data for the entry (i, j), the more we believe that suppressed data in the other years are close to the cut-off level $(H + 1)$ — and most likely equal to the highest suppressed value H. We have implemented this in our objective function via the term:

$$\sum_{(i,j)\in\mathscr{S}_t, t\in\mathscr{T}} \Big(\exp\left[f_{ij}(H - T_{ijt}) \right] - 1 \Big) \tag{2}$$

We will use this term to penalize the objective function. The larger the difference between H and T_{ijt}, the more we penalize the objective function, and the penalty is higher the larger the value on f_{ij}. This favors values of T_{ijt} that are close to H.

An alternative way of handling this is through a transition-based idea: If a suppressed value is either preceded by an unsuppressed value, superseded by an unsuppressed value, or both, a reasonable hypothesis is that the suppressed value is close to H. Let $\mathscr{T}^t$ denote the set of indices where one of the above-quoted properties holds. We can then consider a penalty term on the form

$$\sum_{(i,j,t)\in\mathscr{T}^t} \Big(\exp\left[(H - T_{ijt}) \right] - 1 \Big) \tag{3}$$

Again, this term favors values of T_{ijt} that are close to H, but only if the variable in entry (i, j)

at time t is super-seeded or presided by unsuppressed values. We tried several other ways of implementing penalties, but these proved unsuccessful in improving the accuracy, while the two terms above markedly improved it.

5.2 Combining Time Series and Entropy

When we modify the objective function to include the penalty terms, the question of how they should be weighted arises. We consider the following objective function:

$$\begin{aligned} \max_{T_{ijt}} \quad & -\gamma \sum_{(i,j)\in\mathscr{S}_t} T_{ijt}\cdot\ln(T_{ijt}) \\ & -(1-\gamma)\delta \sum_{(i,j)\in\mathscr{S}_t} \Big(\exp\big[f_{ij}(H-T_{ijt})\big]-1\Big) \\ & -(1-\gamma)(1-\delta) \sum_{(i,j,t)\in\mathscr{T}^t} \Big(\exp\big[(H-T_{ijt})\big]-1\Big) \end{aligned} \tag{4}$$

Here $\gamma \in [0,1]$ and $\delta \in [0,1]$ are hyperparameters determining how much weight we should put on the different terms in the objective function. The case $\gamma = 1$ corresponds to the original problem where all weight is put on entropy. If $\delta = 1$, the transition term has no impact, while the case $\delta = 0$ attributes full impact to the transition term.

As it is difficult to say apriori which values of the hyperparameters offer the best performance, we experimented with many different cases. The results indicated that a cluster of the consistently best results, in terms of accuracy, formed at, and close to, the hyperparameters $(\gamma,\delta) = (0.8, 0.2)$. Hence, $(\gamma,\delta) = (0.8, 0.2)$ are our hyperparameters of choice, giving the entropy, fraction, and transition terms an 80%, 4%, and 16% weight, respectively. However, the accuracy deviated only moderately across alternatives, indicating that the choice of hyperparameters is not critical for the model's performance.

5.3 Results Based on Incorporating Information on Time Series

The results presented in Table 4 on the effect of incorporating information from time series and distances reinforces the conclusion that entropy maximization performs exceptionally well in disclosing information that was suppressed. The improvements in Accuracy and SRMSE are not substantial when either fractions (Equation 2) or transitions (Equation 3 are incorporated into the formulation of the maximization problem, with $\delta = 1$ and $\delta = 0$, respectively. It follows from Table 4 that adjusting for the effect of time series varies somewhat over the years. For the overall 20-year period, however, they each contribute to a reduction in SRMSE from about 0.39 to about 0.37, while Accuracy increases from about 0.75 to about 0.77.

Table 4: Results from disclosing suppressed cell information by a pure entropy maximization approach, and by approaches in addition incorporating time series information.

			Pure		$(\gamma,\delta)=(0.8,1)$		$(\gamma,\delta)=(0.8,0)$		$(\gamma,\delta)=(0.8,0.2)$	
Year	$\lvert CELL\rvert$	$\sum$	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE
1	66	89	0.6515	0.4654	0.7576	0.3651	0.6515	0.4654	0.7273	0.3873
2	67	85	0.7313	0.4086	0.7313	0.4086	0.8806	0.2724	0.8806	0.2724
3	66	80	0.7121	0.4763	0.7879	0.3800	0.7879	0.3800	0.8182	0.3518
4	65	77	0.6923	0.4683	0.6923	0.4683	0.7231	0.4442	0.7231	0.4442
5	69	84	0.6812	0.4638	0.6812	0.4638	0.7101	0.4422	0.7681	0.3956
6	73	93	0.6712	0.4501	0.8082	0.3437	0.8082	0.3437	0.8082	0.3437
7	67	84	0.7313	0.4134	0.7313	0.4134	0.7313	0.4134	0.7313	0.4134
8	64	76	0.8125	0.3646	0.7812	0.3939	0.8125	0.3646	0.8125	0.3646
9	65	78	0.7538	0.4134	0.6923	0.4623	0.7846	0.3867	0.7846	0.3867
10	64	77	0.8125	0.3599	0.7812	0.3887	0.8125	0.3599	0.8125	0.3599
11	68	89	0.7941	0.3467	0.7059	0.4144	0.6765	0.4346	0.7059	0.4144
12	66	88	0.7879	0.3454	0.8788	0.2611	0.6970	0.4129	0.8788	0.2611
13	67	91	0.7612	0.3598	0.7612	0.3598	0.7612	0.3598	0.8209	0.3116
14	67	85	0.7463	0.4307	0.8209	0.3336	0.8060	0.3852	0.8209	0.3336
15	70	94	0.7714	0.3560	0.7714	0.3560	0.7714	0.3560	0.7714	0.3560
16	65	78	0.7538	0.4134	0.7538	0.4134	0.7846	0.3867	0.7846	0.3867
17	68	86	0.8235	0.3322	0.8235	0.3322	0.9412	0.1918	0.9412	0.1918
18	72	97	0.7778	0.3499	0.7500	0.3711	0.7500	0.3711	0.7778	0.3499
19	75	108	0.7867	0.3208	0.9467	0.1604	0.7600	0.3402	0.8667	0.2536
20	76	109	0.8158	0.2993	0.7632	0.3393	0.7895	0.3199	0.8421	0.2771
Total	1360	1748	0.7537	0.3913	0.7721	0.3715	0.7721	0.3750	0.8044	0.3441

In the approach where both fractions and transitions are accounted for, with hyperparameters $(\gamma,\delta)=(0.8,0.2)$, the suppressed information is to a larger degree disclosed, with Accuracy of about 0.80 and SRMSE of about 0.34. Hence, adjusting for time series information improves fit to the actual observations. However, it remains to be seen whether this contribution has a significant practical impact on estimation results and predictions when data are implemented in a spatial interaction model.

Both Accuracy and SRMSE measure to what degree missing data are successfully replicated. An alternative way of presenting the degree of replication is in terms of the matrix in Table 5.

Table 5: The replication matrix corresponding to a procedure with entropy maximization and time series adjustments; $(\gamma, \delta) = (0.8, 0.2)$. Aggregated over all the years.

	0	1	2
0	26	14	0
1	54	719	119
2	0	79	349

This represents a very straightforward way to interpret how the prescribed procedure successfully discloses the suppressed information. Consider, for example, all the origin-destination combinations with just one observed commuter. The procedure correctly predicted 719 of these 892 combinations. Notice also that the prediction never deviates from the actual number by more than one commuter over the entire period.

6 Evaluating Extended Rules of Suppression

The results presented in the previous sections demonstrate that entropy maximization and time series adjustments do very well in replicating suppressed cell information. As pointed out by, for example, Abowd and Schmutte (2019), the availability of linearly dependent statistics, like marginal constraints, represents a potential for reconstructing confidential variables. Hence, privacy concerns may call for some action from the data-releasing agency in cases where such information is available. The agency can choose to suppress information on the marginal sums of the commuting matrix. This option is discussed in Section 6.1, while Section 6.2 addresses the option of raising the cut-off suppression value.

6.1 What if Marginal Totals Are Suppressed?

One option is to suppress the row sums, representing the spatial residential location pattern of the workers. Another option is to suppress the column sums, that is, information on the number of jobs in each of the towns in the geography. Finally, it is, of course, also an option to suppress both row and column sums. We made several approaches to disclose the suppressed information in these cases. Entropy maximization proved to be an essential part of all the approaches that succeeded best in disclosing suppressed information.

The results in Table 6 are based on the scenario where one worker is added to all cells with 0 or 1 commuter in year 0. The results from experiments with time series information are not included in the table, as they did not contribute to replicating the actual commuting matrices over the years.

Table 6: Year-by-year comparisons of disclosing suppressed information in cases where row and column sums are both KNOWN, ROW sums are suppressed, COLUMN sums are suppressed, and BOTH rows and columns sums are suppressed, respectively.

			KNOWN		ROW		COLUMN		BOTH	
Year	$\|CELL\|$	$\sum$	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE	Acc.	SRMSE
1	66	89	0.7273	0.3873	0.6212	0.4830	0.5303	0.5773	0.5455	0.5477
2	67	85	0.8806	0.2724	0.5373	0.6090	0.6269	0.5096	0.5373	0.6531
3	66	80	0.8182	0.3518	0.5606	0.6736	0.5909	0.5562	0.5000	0.6581
4	65	77	0.7231	0.4442	0.4154	0.7404	0.6000	0.5923	0.4769	0.7550
5	69	84	0.7681	0.3956	0.4783	0.6851	0.6957	0.4845	0.4928	0.7400
6	73	93	0.8082	0.3437	0.5753	0.5357	0.6164	0.4861	0.4795	0.6496
7	67	84	0.7313	0.4134	0.5672	0.6007	0.6269	0.5682	0.5821	0.6609
8	64	76	0.8125	0.3646	0.4844	0.7293	0.5312	0.6316	0.4531	0.7443
9	65	78	0.7846	0.3867	0.5385	0.7161	0.5692	0.6027	0.4923	0.7596
10	64	77	0.8125	0.3599	0.5000	0.6892	0.5469	0.5877	0.5000	0.6892
11	68	89	0.7059	0.4144	0.5441	0.5860	0.6029	0.6005	0.5147	0.6419
12	66	88	0.8788	0.2611	0.6667	0.5383	0.6515	0.4707	0.5455	0.5539
13	67	91	0.8209	0.3116	0.5821	0.5245	0.6567	0.4587	0.5672	0.5088
14	67	85	0.8209	0.3336	0.6119	0.5447	0.6716	0.4517	0.5821	0.5615
15	70	94	0.7714	0.3560	0.5857	0.5487	0.6857	0.4710	0.6429	0.5629
16	65	78	0.7846	0.3867	0.6154	0.6027	0.6154	0.6027	0.5077	0.6856
17	68	86	0.9412	0.1918	0.6176	0.4889	0.6471	0.5252	0.6029	0.5252
18	72	97	0.7778	0.3499	0.6667	0.4285	0.6528	0.4629	0.6250	0.4791
19	75	108	0.8667	0.2536	0.6800	0.3928	0.6933	0.4089	0.7467	0.3761
20	76	109	0.8421	0.2771	0.5789	0.4524	0.6579	0.4524	0.6711	0.4232
	1360	1748	0.8044	0.3441	0.5728	0.5739	0.6250	0.5219	0.5559	0.6041

Table 6 refers to 4 different scenarios of available information. The KNOWN scenario is where both row and column sums are known, while row and column sums are suppressed in scenarios ROWS and COLUMNS, respectively. In BOTH scenarios, both information on row sums and column sums are suppressed. In all the cases in Table 6, assume that we still have reliable information on the total number of workers and jobs in the geography, that is, the sum of all rows and columns, respectively. If so, we have information on the total number of suppressed workers but not how they are distributed between cells, rows, and columns, in the case where both row and column sums are suppressed.

The scenario BOTH corresponds to the following optimization problem:

$$\max_{T_{ijt}} - \sum_{i,j \in I, t \in \mathscr{T}} T_{ijt} \log(T_{ijt})$$

under the constraint

$$\sum_{i,j\in I} T_{ijt} = \sum_{i,j\in I} T_{ijt}^{\text{non-suppressed}} \qquad \text{(assuming that the sum on the right is public)}$$

and where

$$T_{ijt} = \begin{cases} T_{ijt}^{\text{non-suppressed}} & (i,j) \notin \mathscr{S}_t \\ \text{a free optimization variable} & (i,j) \in \mathscr{S}_t \end{cases}$$

The scenarios ROWS and COLUMNS are defined similarly. The results are relatively encouraging from the position of the data releasing agency. It follows from Table 6 that suppressing information on both marginal totals in rows and columns leads to a reduction of the Accuracy from a level of around 0.80 to a level of about 0.55, and SRMSE increases from about 0.34 to about 0.60. However, in considering these figures, remember that they refer to only the suppressed cells of the commuting matrix. This means that the deviations between the replicated and the actual total matrices will be relatively minor also in the case where row and column sums are suppressed. Hence, it seems reasonable to hypothesize that these deviations at least lead to a less severe bias in estimating parameters representing commuting behavior than in cases where the suppressed information is ignored.

Another interesting result from Table 6 is that the potential to disclose the suppressed cells is not substantially increased if only row or column sums are suppressed. This, in particular, applies for the case where row sums are suppressed, on average resulting in an Accuracy of around 0.57. This means that the statistical agency is recommended to suppress row sums, while privacy is not harmfully reduced when data on column sums are released. Suppressing just column sums leads to an Accuracy of around 0.63. On the other hand, information on the spatial distribution of jobs may be helpful for several research projects.

Why is it that suppressing row sums is more challenging from a privacy concern than suppressing column sums? Technically, it can be seen from the matrix of commuting flows in Table 1 that all the rows have specific cells with some commuters below the cut-off value of 3. At the same time, this is the case for 11 of the 12 columns of the matrix. Hence, the information loss of suppression is somewhat lower for the columns. There are reasons to think this is not just a coincidence specific to this data set. The observation that one column has large numbers reflects the tendency of firms and jobs to cluster in, for example, a regional center, that is, town E in our geography. This is due to urbanization economies and Marshallian agglomeration forces, giving rise to economies of localization, resulting in a location pattern where jobs are spatially more concentrated than the residential location pattern.

6.2 Experimenting With the Cut-Off Suppression Value

The discussion above demonstrated that suppressing marginal totals may not be sufficient to limit the disclosure of sensitive information. Another action from the statistical agency may be to raise the cut-off value from suppressing cell information. Table 9 provides results from experimenting with the cut-off value. The Accuracy is found to fall considerably with increasing the cut-off value above 3. Correspondingly, the SRMSE is found to increase, but not to the same degree. According to our experiments, however, there is a low difference in Accuracy and SRMSE for cut-off values of 4 and 5. This reflects the nature of our data, representing a relatively small population with only 12 zones and a modest variation in the number of low-valued cells in the matrix of commuting flows. Also, remember that these values refer to only the part of the commuting matrix that is suppressed. Hence, an approach resulting in a value of SRMSE around 0.65 represents a substantial improvement compared to a procedure where values are set to zero in all suppressed cells. Even with a cut-off value of 5 observations in a cell, the entropy-based approach of disclosing suppressed information may be removing a substantial source of bias in estimating parameters in a spatial interaction model.

Table 7: Results based on various cut-off values.

			$(\gamma, \delta) = (0.8, 0.2)^s$	
Suppression level	$\|CELL\|$	$\sum$	Acc.	SRMSE
<3	1360	1748	0.5728	0.5739
<4	1561	2351	0.4478	0.6958
<5	1653	2719	0.4459	0.6538

7 Implementing Distance Information

It makes intuitive sense that a short distance increases the probability that the suppressed value is close to H, ceteris paribus. This is according to the highly frequently cited Tobler's first law of geography, which states that "everything is related to everything else, but near things are more connected than distant things" (Tobler, 1970). In the experiments discussed in this section, we will see that the impact of distances depends on what information is available about the marginal totals. It also matters somewhat how distances are defined in the model. The following two alternatives were considered:

globally scaled distances; meaning that all distances are scaled relative to the longest observed distance of all the (i, j)-combinations in the sample

locally scaled distances; meaning that all distances are scaled relative to the longest distance observed for a specific origin $.i$

If a town has a highly accessible location, with many job opportunities in surrounding areas, its inhabitants may not consider long-distance commuting. The labor market accessibility, reflecting the distances to relevant job opportunities, may vary systematically across the towns. Thus, the hypothesis is that locally scaled distances capture relevant characteristics of the spatial structure and potentially contribute to explaining both the observed commuting pattern in the region and the spatial pattern of suppressed information. This hypothesis was supported in our experiments; locally scaled distances outperform globally scaled distances in replicating the actual commuting matrix.

In a standard spatial interaction model, commuting is exponentially deterred by distance. Hence, if the distance between two towns is considerable, it appears more likely that the suppressed value is small than when the distance is small. As Figure 1 illustrates, there are substantial differences in distances between towns. As an example of how this could be implemented, we suggest the following:

$$\begin{aligned} \max_{T_{ijt}} \quad & -\gamma \sum_{(i,j)\in\mathscr{S}_t} T_{ijt}\cdot\ln(T_{ijt}) \\ & -(1-\gamma)\delta \sum_{(i,j)\in\mathscr{S}_t} \Big(\exp\big[f_{ij}(H-T_{ijt})\big]-1\Big) \\ & -(1-\gamma)(1-\delta) \sum_{(i,j,t)\in\mathscr{T}^t} \Big(\exp\big[(1-d_{ij}^G)(H-T_{ijt})\big]-1\Big) \end{aligned} \tag{5}$$

where d_{ij}^G denotes geographical distance normalized to one by division by the largest distance in the system. The effect of the term $(1-d_{ij}^G)$ is that a small T_{ijt} leads to less penalty when the distance between i and j is considerable than when it is small.

7.1 Information on Marginal Totals Is Provided

Introducing locally scaled distances in the transition term, as formulated by Equation 5, adds nothing to disclose the suppressed data. The results are more or less identical to those corresponding to the approach where fractions and transitions are accounted for, with $(\gamma,\delta)=(0.8,0.2)$, with no care taken to distances. We have experimented with several other ways of implementing distances into the objective function than what is represented by Equation 5. However, the conclusion remains the same; none of these implementations led to significantly better performance. Implementing distances in the objective function did not significantly lead to disclosing suppressed values in the cases where the marginal sums are known.

Still, this does, of course, not on a general basis mean that distances should not be used in disclosing suppressed data in a scenario where the marginal totals are known. Such an approach is based on an appealing and theoretically sound hypothesis. Our results may reflect that

data are generated from a small geography with a low variation of values representing long-distance commuting. We cannot rule out the possibility that adjusting for distances contributes substantially to disclosing suppressed information in other test regimes, for example, based on an actual data set from an extensive system of towns and cities. Further exploration of these issues is left for future research.

7.2 Information on Marginal Totals Is Suppressed

Compared to the scenario where the sums of rows and columns are known, the evaluation of adjusting for distances changes substantially in cases where marginal totals are suppressed. Table 8 provides results for a case where row sums are suppressed. In comparing these results to the last two columns of Table 6, it follows that distances employ a noticeable impact in disclosing the correct number of workers in the suppressed cells. The accuracy increases from around 0.56 to around 0.71 due to distance adjustments. Correspondingly, the SRMSE is reduced from around 0.60 to around 0.43. While suppressing the marginal sums decreases the accuracy efficiently, 60.6% of this reduction is recouped by implementing a local distance parameter to the objective function's entropy term ($(\gamma, \delta) = (0.8, 0.2)^{s,d}$). This supports the reasonable hypothesis that distances generate more order and better fit in cases with sparse information on suppressed information patterns.

Table 8: Accounting for locally scaled distances in a case with known marginal sums and in a case with suppressed row sums; $(\gamma, \delta) = (0.8, 0.2)^{s,d}$.

Year	$\lvert CELL \rvert$	$\sum$	Acc.	SRMSE
1	66	89	0.5909	0.5000
2	67	85	0.6716	0.4517
3	66	80	0.7727	0.4308
4	65	77	0.6769	0.5129
5	69	84	0.7246	0.4638
6	73	93	0.6712	0.4501
7	67	84	0.7313	0.4134
8	64	76	0.7344	0.4708
9	65	78	0.7077	0.4848
10	64	77	0.7500	0.4156
11	68	89	0.7059	0.4144
12	66	88	0.6818	0.4523
13	67	91	0.6418	0.4407
14	67	85	0.7015	0.4307
15	70	94	0.7143	0.4538
16	65	78	0.7385	0.4623
17	68	86	0.7941	0.3588
18	72	97	0.7500	0.3711
19	75	108	0.7333	0.3586
20	76	109	0.7632	0.3393
	1360	1748	0.7132	0.4324

As demonstrated in Section 5.3, incorporating time series information improves disclosing suppressed information in cases where the marginal sums are known. The results in Section 6.1, on the other hand, demonstrated that time series information did not significantly add to the performance in cases where marginal totals are suppressed. For distances, the results are the opposite. Information on distances adds substantially to the potential of disclosing suppressed information in instances where row and/or column sums are not known. We hypothesize that the hard marginal total constraints limit the feasible space to such an extent that the soft constraints from distance information do not provide any added value beyond what the time series considerations provide.

The discussion in Section 6.2 demonstrated that raising the cut-off value of cell suppression represents an efficient action to reduce the accuracy and preserve privacy in providing data. This can, of course, also be done in cases where the marginal sums are suppressed. By comparing the results in Table 9 to the results in Table 7, it once again follows that adjusting for information on distances leads to a considerable increase in the Accuracy and a corresponding reduction in the

SRMSE.

Table 9: Results based on various cut-off values in a case with suppressed row and column sums.

			$(\gamma,\delta)=(0.8,0.2)^{s,d}$	
Suppression level	$\lvert CELL\rvert$	$\sum$	Acc.	SRMSE
<3	1360	1748	0.7132	0.4324
<4	1561	2351	0.6028	0.4590
<5	1653	2719	0.5590	0.4662

The replication matrix in Table 10 refers to a cut-off value of 5. This matrix gives a clear perception that entropy-based allocation of suppressed information is superior to assigning a value of 0 in the relevant cells. Specifically, while the accuracy remains moderately high at 0.5590, only around 4.6% of the 1653 predictions have a deviation greater than one from the true values. Just 3 of the 1653 predictions deviate by more than two from the true values. This justifies a hypothesis that entropy-based allocation of suppressed information removes a potentially serious bias in estimating parameters representing spatial labor market mobility. This hypothesis will be examined in the section to follow.

Table 10: The replication matrix corresponding to a procedure with entropy maximization and distance adjustments in a case with a cut-off value of 5. Aggregated over all the years.

	0	1	2	3	4
0	4	23	6	0	0
1	60	579	127	11	0
2	13	193	260	123	25
3	3	17	74	61	47
4	0	0	1	6	20

8 Estimation Results Based on the Doubly Constrained Gravity Model

As pointed out in Section 4.1, the modified maximum entropy problem has as its solution a multinomial logit model. In Equation 1 this model formulation was given by $T_{ij} = A_i B_j e^{-\beta c_{ij}}$, where $A_1,...,A_n, B_1,...,B_n$ are constants and c_{ij} a measure of spatial separation between an origin i and a destination j, defined by the generalized cost of traveling from origin i to destination j. The doubly constrained gravity model is an alternative, equivalent formulation of this trip distribution model. In a commuting context, let O_i represent the number of commuters originating from town i, while D_j is the observed number of trips with destination in zone

j. A_i and B_j are the balancing factors ensuring that $\sum_j T_{ij} = O_i$ and $\sum_i T_{ij} = D_j$. A standard formulation of a doubly constrained gravity model then is:

$$T_{ij} = A_i O_i B_j D_j e^{-\beta c_{ij}} \tag{6}$$

$$A_i = \left[\sum_j B_j D_j e^{-\beta c_{ij}} \right]^{-1} \tag{7}$$

$$B_j = \left[\sum_i A_i O_i e^{-\beta c_{ij}} \right]^{-1} \tag{8}$$

For more details on the theoretical foundation of this model, see for instance Erlander and Stewart (1990) or Sen and Smith (1995). O_i and D_j represent characteristics of the origin and destination towns expected to influence the volume of work-related trips for the specific combination of towns. It is reasonable, however, that the traffic volume of a specific origin-destination combination is also affected by the spatial characteristics of the other towns in the geography. As pointed out by, for example, Persyn and Torfs (2016), this is accounted for by the balancing factors. Still, it can, in general, be argued that a spatial interaction model should, in addition, account for spatial characteristics that are not captured by the balancing factors. This can, for example, be done in terms of the competing destinations model, where a measure of accessibility is included to account for the location relative to competing destinations (Fotheringham (1983) and Pellegrini and Fotheringham (2002)). Gitlesen and Thorsen (2000) provide an interpretation of the competing destinations model in a commuting context and demonstrate that this model is also relevant when commuting is the relevant form of spatial interaction.

It is well known in the literature that leaving out relevant characteristics of spatial structure introduces a source of estimation bias of spatial interaction parameters, see, for instance, Tiefelsdorf (2003). Suppressed data is another source of biased statistical inference, see for instance Carpenter et al. (2022) and Abowd and Schmutte (2016) for a general discussion. This paper considers potential bias resulting from suppressed information on commuting. The second column of Table 11 provides an estimate of the distance deterrence parameter β following from the standard doubly constrained gravity model in a case where no data are suppressed, that is, for the true matrix of commuting trips. The standard errors of the parameter estimates from the true matrix are estimated by bootstrapping and provided in the third column.

The fourth column of Table 11 provides an estimate of the distance deterrence parameter in the case with suppressed information, with a value of 0 in cells with less than $H+1$ commuters.

The estimates do not deviate much from the estimates based on the non-suppressed matrix of commuting flows. However, as illustrated in Figure 2, there is a systematic difference over the years. Ignoring the suppressed information leads to consistently higher estimates of the distance deterrence parameter. As mentioned in Section 2, Carpenter et al. (2022) claim that bias may follow from the reasonable possibility that less populous geographies have a higher share of suppressed cells. Similarly, in commuting, suppressed cells may mainly come about for long-distance journeys to work.

The differences between the estimates from the non-suppressed and the suppressed matrices of commuting trips are not significantly different from 0. A hypothesis that the estimates based on the two matrices are equal cannot be rejected for any year in the period; the confidence interval of all the estimated differences incorporates 0. When evaluating our approach's performance, we should keep in mind that there are two different sources of uncertainty in these matrices. One source is purely random and is caused by the actions of a finite number of agents. The other source is systematic and is caused by the suppression of data. Suppression of data will, in most cases, lead to a lower number of long-distance commuters, leading to systematically lower estimates for the distance deterrence parameter. A central finding in this paper is that our methods reduce the systematic bias to such small numbers that it is effectively zero.

When the total population increases, the uncertainty due to random choices goes down. In an artificial experiment, we split all the zones into two sub-zones. In the corresponding 24×24 trip-distribution, we allocated the numbers in the original 12×12 trip-distribution to each 2×2 block. The total number of agents in the system increased by four. As expected, the standard deviation in the parameter estimates was roughly 50% of the original values. The optimal parameter values changed, but when we compared the non-suppressed and suppressed matrices, the systematic bias slightly increased. This hints that the systematic bias may dominate the random bias in a system with many agents distributed across a geography with many zones.

The second last column of Table 11 provide estimates of the distance deterrence parameter based on commuting matrices where the suppressed information is disclosed by the methods explained in the previous sections. These estimates refer to a data matrix where the marginal totals are known, and they almost precisely match the corresponding estimates for the non-suppressed trip distribution. To a slightly lower degree, this is also the case for the data matrix where marginal totals are suppressed. The suppressed information is next disclosed by the procedure where the entropy-based method is supplemented by locally scaled distances. The results based on this case are not presented in Table 11, but they are illustrated in Figure 2. The curves marked in red and blue overlap for the major parts and can only be distinguished from one another in a few segments. Hence, from a practical point of view, the bias resulting from suppressed information is also removed in this case.

We have also experimented by varying the cut-off value for suppressing information in the cells of the OD-matrix. As expected, higher cut-off values result in higher estimates of the distance deterrence parameter. However, without entering into details, the increase is not found to be very sensitive to the chosen cut-off value. The reason for this can be found by considering the geography and the synthetic population underlying the experiments. Due to the considerable computing time involved, we kept our agent-based population relatively small, distributed across only 12 towns. As illustrated in Figure 1, the towns are organized into three clusters. The distances are long between the three clusters but short between the towns within each cluster. This explains a pattern of commuting flows with just a few commuters between towns belonging to different clusters but considerable commuting between towns within the same cluster. See the matrix of commuting flows in Table 1. In many, more densely populated geographies, a larger continuum can be expected in both sizes and distances across towns and cities. The commuting pattern will reflect more heterogeneity in sizes and distances, with more continuous variation expected in the number of commuters between the different origin-destination combinations. In such a case, the $\hat{\beta}$'s are expected to be considerably more sensitive to the cut-off value of cell suppression. Hence, cell suppression represents a source of seriously biased estimates and predictions. However, entropy-based approaches to disclosing suppressed information are found to have the potential to eliminate this source of bias.

Our experiments are run for a relatively sparsely populated geography. It is a reasonable hypothesis that the estimation bias resulting from suppression is more severe and significant for real geographies with more towns, more workers, and more substantial variation in distances, particularly between towns far from each other. We leave the testing of such a hypothesis for future research.

Table 11: Estimates of the distance deterrence parameter β in a standard doubly-constrained gravity model. Estimates based on an non-suppressed commuting pattern, a suppressed commuting pattern ($H+1=3$), and by data where suppressed cells are disclosed by time-series adjusted entropy maximization.

	Non-suppressed		Suppressed		$(\gamma,\delta)=(0.8,0.2)^{s,d}$	
Year	Mean	Std. Err.	Mean	Std. Err.	Mean	Std. Err.
2000	0.0659	0.00231	0.0667	0.00194	0.0659	0.00212
2001	0.0656	0.00198	0.0663	0.00222	0.0655	0.00211
2002	0.0659	0.00229	0.0665	0.00237	0.0658	0.00222
2003	0.0656	0.00198	0.0661	0.00217	0.0655	0.00191
2004	0.0633	0.00207	0.0639	0.00197	0.0633	0.00229
2005	0.0625	0.00213	0.0633	0.00212	0.0625	0.00211
2006	0.0629	0.00214	0.0636	0.00227	0.0630	0.00224
2007	0.0622	0.00212	0.0627	0.00234	0.0621	0.00204
2008	0.0610	0.00200	0.0616	0.00206	0.0610	0.00219
2009	0.0599	0.00205	0.0604	0.00205	0.0599	0.00206
2010	0.0596	0.00198	0.0600	0.00208	0.0596	0.00199
2011	0.0590	0.00197	0.0595	0.00204	0.0590	0.00187
2012	0.0575	0.00209	0.0580	0.00197	0.0575	0.00187
2013	0.0573	0.00198	0.0578	0.00189	0.0573	0.00191
2014	0.0585	0.00193	0.0591	0.00217	0.0585	0.00201
2015	0.0586	0.00206	0.0590	0.00197	0.0586	0.00196
2016	0.0588	0.00201	0.0592	0.00217	0.0588	0.00192
2017	0.0591	0.00196	0.0596	0.00213	0.0591	0.00203
2018	0.0604	0.00193	0.0609	0.00197	0.0603	0.00192
2019	0.0588	0.00215	0.0595	0.00212	0.0588	0.00188
Average	0.0611	0.00206	0.0617	0.00210	0.0611	0.00203

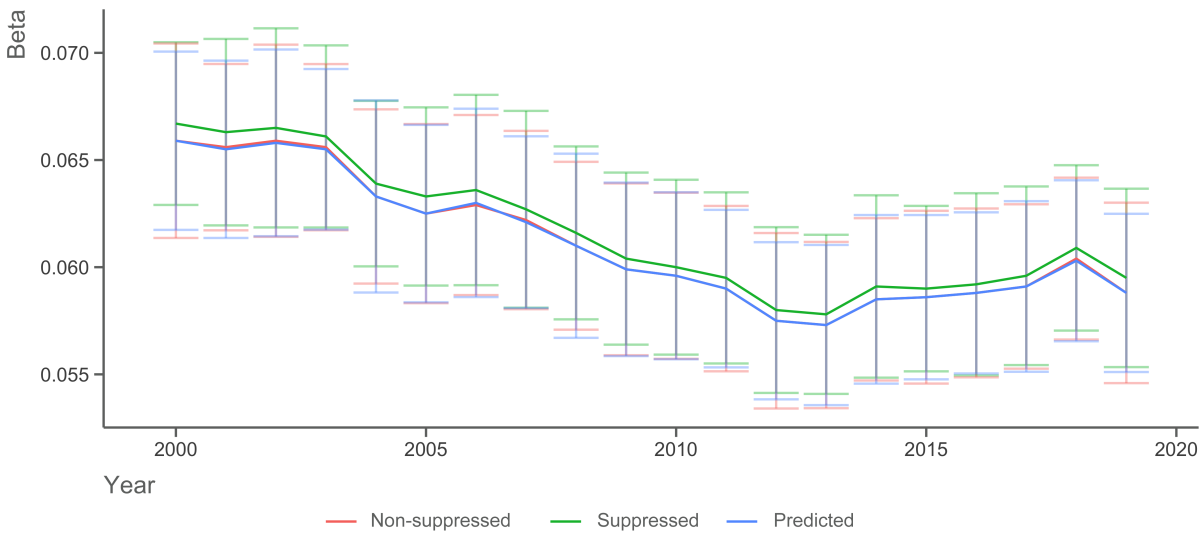

Figure 2: $\hat{\beta}$ values for non-suppressed, suppressed, and predicted $\left((\gamma,\delta)=(0.8,0.2)^{s,d}\right)$ data sets.

9 Summary and Concluding Remarks

There has been a growing concern for privacy issues recently, reinforced by the increasing access to microdata. This has been inducing statistical agencies to introduce statistical disclosure limitations in releasing data. The most commonly used method for tabular data is cell suppression. This represents a challenge for empirical research (Abowd & Schmutte, 2019), representing a source of measurement error that may lead to biased parameter estimates and/or increased standard errors (Carpenter et al., 2022). Hence, disclosure limitations come at a price regarding a potential utility loss in research. In this paper, we discuss how this utility loss can be reduced through attempts to disclose the information represented by the suppressed cells.

Journeys-to-work maybe not in itself represent a very sensitive type of information. However, it has the potential of being linked to external data sources that can be subject to malicious use by intruders. Our discussion is conducted in terms of a case where the spatial dimension leads to cells in an origin-destination matrix with just a few commuters, facilitating identity disclosures. In addition, our results are generally valid for other, more sensitive, social matters than commuting.

Based on synthetic data mirroring the Norwegian population, we demonstrate that a constrained entropy maximizing approach to a large degree succeeds in disclosing the information hidden by suppressing information in cells with less than three commuters. The constraints represent the marginal sums in the matrix of commuting trips, defining the given and known number of workers living and working in each zone of the geography. We also demonstrate that

incorporating information on time series in an entropy-maximizing approach adds significantly to disclosing suppressed information. In contrast, data on distances does not turn out to be relevant in the case with a cut-off suppression value of 3.

As pointed out by Abowd and Schmutte (2016), statistical limitation disclosure is ignorable if correct inferences are made without explicitly accounting for SDL, but (Carpenter et al., 2022) in general call for greater awareness in documenting the methods used in estimation based on suppressed cell data sets. We provide estimates of the distance deterrence parameter in a standard doubly constrained gravity model for data sets based on different treatments of suppressed cells. Ignoring the information in the suppressed cells is found to cause a bias, as expected, with an overvalued estimate of the distance deterrence in commuting. This bias is, to a large degree, eliminated when entropy-based approaches disclose the suppressed information. Quantitatively, these effects are not found to be strong for our synthetic data, but ignoring suppressed information is demonstrated to be potentially harmful, and the results are encouraging in the sense that they prove potential for entropy-maximizing approaches to avoid biased estimates and predictions.

It is well known in the literature that if many linearly dependent statistics are available, there is a substantial potential for a reconstruction attack, a reconstruction of confidential variables, which is a data break (see, for instance, Abowd and Schmutte, 2019). The marginal sums in the matrix of commuting flows are examples of this kind of statistics. What if statistical agencies undertake a secondary suppression to avoid such attacks, that is, to suppress information on the marginal sums? According to our results, leaving out information on marginal totals does not lead to a substantial limitation of the possibility of disclosing suppressed information about commuting trips. An alternative step to preserve privacy is that the statistical agency chooses to increase the cut-off value of cell suppression. Still, our results demonstrate that an entropy-based procedure can remove a substantial source of bias in cases with a higher cut-off value of cell suppression.

Another problem we have addressed is the potential of integrating additional information on the system into the entropy maximizing procedure, discussing the degree to which this contributes to disclosing the suppressed information. In many cases, information on commuting trips is available for several successive periods, like the 20 years period that we consider. In addition, information is generally available on distances between alternative origin-destination combinations of journeys to work. As mentioned above, information on time series is useful, while data on distance are not in a case with known marginal totals. This conclusion is reversed when we have no information on marginal totals. Distances contribute to significantly better fit in such a case with sparse information on the pattern of suppressed information, while time series information does no longer contribute. Based on our synthetic data experiments, we also

found that the potential of an identity disclosure is not substantially increased if only row or column sums are suppressed.

The waterfall chart in Figure 3 summarizes some of the results of different approaches to disclosing suppressed information. First, a pure entropy maximization restores 75.37% of the loss in accuracy that follows from suppressing all cells with less than three observations. Accounting for time series information contributes to restoring another 5.07% of the initial loss in accuracy. However, suppressing marginal totals and raising the cut-off value reduces the accuracy considerably, but accounting for locally scaled distances between towns restores a significant part of this loss. The accuracy restoration is found to more or less remove the estimation bias of the distance deterrence parameter in a doubly constrained gravity model.

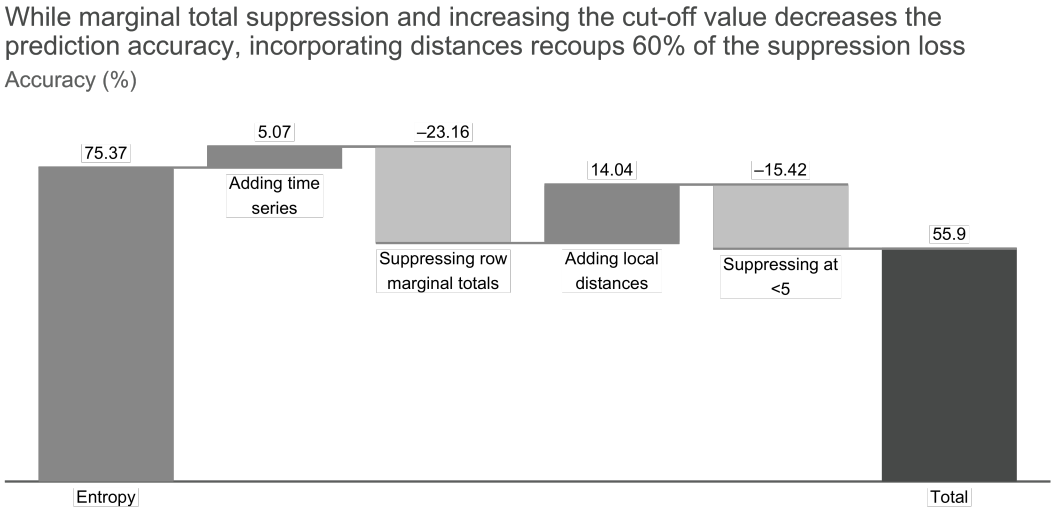

Figure 3

The literature distinguishes between identity disclosure and inferential disclosure, where the former can be argued to be the most important kind of disclosure to avoid (G. Duncan and Lambert, 1989, Heldal and Fosen, 2001, Airoldi et al., 2011, and Abowd and Schmutte, 2016). As stated in the introduction, Abowd and Schmutte (2016) claimed that it is, from a probabilistic perspective, impossible to release data without compromising confidentiality. They also claim that there should be less trepidation in using data that have been the subject of statistical disclosure limitation; in general, it is not more serious than other sources of non-ignorable missing data.

Abowd and Schmutte (2019) recommended that adding noise to the micro-data may be a better alternative to suppression, and Abowd and Schmutte (2016) claimed that data-releasing agencies

are becoming more open to the use of noise-inferred methods in producing data tables. This is useful in reaching unbiased parameters, as opposed to what is in general following from approaches with count-based suppression. On the other hand, estimates based on noise infusion are less precise, which according to Chetty and Friedman (2019), is a crucial drawback and a primary concern in much research. As a general recommendation, statistical agencies should be encouraged to use methods that, to a low degree, limit the statistical validity of the study and, to a high degree, makes it possible to reach results that correspond to the results that would follow from actual data, with no measurement errors.

The main contribution of this paper is twofold. First, we demonstrate that suppressing the information in cells with a low number of observations does not necessarily preserve privacy adequately. Such an approach from the data releasing agency is opening for identity disclosure. The agencies should consider suppressing marginal totals, using higher cut-off values, and/or other methods, such as noise infusion. Second, our results strongly recommend that researchers develop sound methods to adjust for suppressed information rather than just ignoring the cells by setting the values equal to zero. Well-founded methods to adjust for suppressed information can potentially remove a source of harmful bias in estimating basic parameters and prevent invalid predictions, for instance, on important policy issues. Our study is performed from a spatial labor market interaction perspective, but it is at least a reasonable hypothesis that these conclusions are valid also from a more general perspective.

References

Abowd, J. M., & Schmutte, I. M. (2016). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, *2015*(1), 221–293.

Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, *109*(1), 171–202.

Airoldi, E. M., Bai, X., & Malin, B. A. (2011). An entropy approach to disclosure risk assessment: Lessons from real applications and simulated domains. *Decision Support Systems*, *51*(1), 10–20.

Anas, A. (1983a). Cities and complexity: Understanding cities through cellular automata, agent based models and fractals. *Transportation Research Part B: Methodological*, *17*(1), 13–23.

Anas, A. (1983b). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, *17*(1), 13–23.

Bartik, T. J., Biddle, S. C., Hershbein, B. J., & Sotherland, N. D. (2018). Wholedata: Unsuppressed county business patterns data: Version 1.0 [dataset]. *Kalamazoo: WE Upjohn Institute for Employment Research*.

Bezzi, M. (2007). An entropy based method for measuring anonymity. *2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops-SecureComm 2007*, 28–32.

Binswanger, J., & Oechslin, M. (2020). Better statistics, better economic policies? *European Economic Review*, *130*, 103588.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 200–217.

Carpenter, C. W., Van Sandt, A., & Loveridge, S. (2022). Measurement error in us regional economic data. *Journal of Regional Science*, *62*(1), 57–80.

Chetty, R., & Friedman, J. N. (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples. *AEA Papers and Proceedings*, *109*, 414–20.

Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, *6*(1), 73–85.

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, *7*(2), 207–217.

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). Why statistical confidentiality? *Statistical confidentiality* (pp. 1–26). Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284.

Erlander, S. (2010). *Cost-minimizing choice behavior in transportation planning: A theoretical framework for logit models*. Springer Science & Business Media.

Erlander, S., & Stewart, N. F. (1990). *The gravity model in transportation analysis: Theory and extensions* (Vol. 3). Vsp.

Fotheringham, A. S. (1983). A new set of spatial-interaction models: The theory of competing destinations. *Environment and Planning A: Economy and Space*, *15*(1), 15–36.

Gholami, A., Thorsen, I., & Ubøe, J. (2022). *An agent-based approach to study spatial structure effects on estimated distance deterrence in commuting*. NHH working paper.

Gitlesen, J. P., & Thorsen, I. (2000). A competing destinations approach to modeling commuting flows: A theoretical interpretation and an empirical application of the model. *Environment and Planning A*, *32*(11), 2057–2074.

Gjestland, A., Thorsen, I., & Ubøe, J. (2006). Some aspects of the intraregional spatial distribution of local sector activities. *The Annals of Regional Science*, *40*(3), 559–582.

González-Vidal, A., Rathore, P., Rao, A. S., Mendoza-Bernal, J., Palaniswami, M., & Skarmeta-Gómez, A. F. (2020). Missing data imputation with bayesian maximum entropy for internet of things applications. *IEEE Internet of Things Journal*, *8*(21), 16108–16120.

Guldmann, J.-M. (2013). Analytical strategies for estimating suppressed and missing data in large regional and local employment, population, and transportation databases. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, *3*(4), 280–289.

Heldal, J., & Fosen, J. (2001). *Statistisk Konfidensialitet i SSB* (Vol. 28). Statistics Norway.

Irwin, E. G. (2010). New directions for urban economic models of land use change: Incorporating spatial dynamics and heterogeneity. *Journal of Regional Science*, *50*(1), 65–91.

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, *110*(9), 2819–58.

Li, C., Miklau, G., Hay, M., McGregor, A., & Rastogi, V. (2015). The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB Journal*, *24*(6), 757–781.

Majeed, A., & Lee, S. (2020). Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data. *Applied Intelligence*, *50*(8), 2555–2574.

Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, *105*(1), 272–98.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1–29.

Mattson, L.-G., & Weibull, J. (2002). Probabilistic choice and procedurally rationality. *Games and Economic Behavior*, *41*(1), 61–78.

McArthur, D. P., Thorsen, I., & Ubøe, J. (2012). Labour market effects in assessing the costs and benefits of road pricing. *Transportation Research Part A: Policy and Practice*, *46*(2), 310–321.

McArthur, D. P., Thorsen, I., & Ubøe, J. (2010). A micro-simulation approach to modelling spatial unemployment disparities. *Growth and Change*, *41*(3), 374–402.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, *3*(4), 303–328.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, *6*(4), 487–500.

Page, S. E. (1999). On the emergence of cities. *Journal of Urban Economics*, *45*(1), 184–208.

Pellegrini, P. A., & Fotheringham, A. S. (2002). Modelling spatial choice: A review and synthesis in a migration context. *Progress in Human Geography*, *26*(4), 487–510.

Persyn, D., & Torfs, W. (2016). A gravity equation for commuting with an application to estimating regional border effects in belgium. *Journal of Economic Geography*, *16*(1), 155–175.

Rajasekaran, S., Harel, O., Zuba, M., Matthews, G., & Aseltine, R. (2009). Responsible data releases. *Industrial Conference on Data Mining*, 388–400.

Rodrigues, J. D. (2016). Maximum-entropy prior uncertainty and correlation of statistical economic data. *Journal of Business & Economic Statistics*, *34*(3), 357–367.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Sen, A., & Smith, T. E. (1995). Gravity models: An overview. *Gravity Models of Spatial Interaction Behavior*, 49–152.

Shlomo, N. (2018). Statistical disclosure limitation: New directions and challenges. *Journal of Privacy and Confidentiality*, *8*(1).

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Tiefelsdorf, M. (2003). Misspecifications in interaction model distance decay relations: A spatial structure effect. *Journal of Geographical Systems*, *5*(1), 25–50.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*(sup1), 234–240.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge University Press.

Tsiafoulis, S. G., Zorkadis, V. C., & Pimenidis, E. (2011). Maximum entropy oriented anonymization algorithm for privacy preserving data mining. *Global security, safety and sustainability & e-democracy* (pp. 9–16). Springer.

Willenborg, L., & De Waal, T. (2012). *Elements of statistical disclosure control* (Vol. 155). Springer Science & Business Media.

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*(3), 253–269.

Wilson, A. (2010). Entropy in urban and regional modelling: Retrospect and prospect. *Geographical Analysis*, *42*(4), 364–394.

NORGES HANDELSHØYSKOLE
Norwegian School of Economics

Helleveien 30
NO-5045 Bergen
Norway

T +47 55 95 90 00
E nhh.postmottak@nhh.no
W www.nhh.no