Freedom of Information and Personal Confidentiality in Spatial Covid-19 Data

Michael Beenstock, Department of Economics, Hebrew University of Jerusalem

Daniel Felsenstein*, Department of Geography, Hebrew University of Jerusalem

*(corresponding author: daniel.felsenstein@mail.huji.ac.il)

Abstract

We draw attention to how, in the name of protecting the confidentiality of personal data, national statistical agencies have limited public access to spatial data on Covid-19. We also draw attention to large disparities in the way that access has been limited. In doing so, we distinguish between absolute confidentiality in which the probability of detection is 1, relative confidentiality where this probability is less than 1, and collective confidentiality, which refers to the probability of detection of at least one person. In spatial data the probability of personal detection is less than 1, and the probability of collective detection varies directly with this probability and Covid-19 morbidity. Statistical agencies have been concerned with relative and collective confidentiality, which they implement using the techniques of truncation, where spatial data are not made public for zones with small populations, and censoring, where exact data are not made public for zones where morbidity is small.

Granular spatial data are essential for epidemiological research into Covid-19. We argue that in their reluctance to make these data available to the public, data security officers (DSO) have unreasonably prioritized data protection over freedom of information. We also argue that by attaching importance to relative and collective confidentiality, they have over-indulged in data truncation and censoring. We highlight the need for legislation concerning relative and collective confidentiality, and regulation of DSO practices regarding data truncation and censoring.

Acknowledgements: Many thanks to Tal Zarsky,Guy Zomer and the reviewers for their comments.

1. Introduction

As in so many areas, the Covid-19 pandemic has imposed a sea-change on government statistical agencies. In their quest to track, contain and forecast the spread of the virus, governments have been forced to address new data governance and privacy challenges (OECD 2020a). While many of these are related to the nature of digital data sources such as mobile phone data and biometrics (Newlands et al 2020), demands are also being made on more traditional statistical sources such as censuses, household and income surveys and tax data. In the case of Covid-19 data, these demands call for a far from perfect trade-off between data accessibility and freedom of information for containing the pandemic on the one hand and issues of personal confidentiality on the other (OECD 2020b).

In this paper we address this trade-off in the context of spatial Covid-19 data. Since the outbreak of the Covid-19 pandemic national statistical agencies (NSA's) have been inundated with data requests from scientific investigators, the media and organizations concerned with public health. Much of this demand has been for spatial data where information on individuals is aggregated into territorial units or zones of differing levels of resolution. As both the transmission of Covid-19 and the policy response to its spread are inherently spatial (Poom et al 2020), government agencies are increasingly requested to supply data to track and analyze the spatiotemporal dynamics of the pandemic. Given this brief, statistical agencies find themselves caught between the hammer of freedom of information and the anvil of protecting individual confidentiality.

Because spatial data are aggregated into zones, this tension should ostensibly be mediated. Reporting, for example, the number of people infected in a zone does not reveal the identity of individuals. We argue, however, that statistical agencies have confounded the absolute confidentiality of personal information, which is the objective of existing legislation, with broader concepts of confidentiality not covered by existing legislation. These include relative confidentiality, which is concerned with the probability of identification faced by each individual, and collective confidentiality, which is concerned with the probability faced by statistical agencies that at least one individual will be identified. These concepts are developed further below.

We argue that these broader concepts of confidentiality have been applied by statistical agencies, such as ministries of health, to limit public access to spatial Covid-19 data. Since these broader concepts are not covered by existing legislation, freedom of information has been unnecessarily infringed.

We draw specific attention to spatial data for morbidity, hospitalizations and mortality in zones such as cities, towns, administrative districts, neighborhoods, census tracts and postal zip codes. These zones are particularly important for mitigation policy and research because Covid-19 is contagious, and its transmission is fundamentally spatial. They are also important more generally, because the public has the right to know for their own safety where the disease is particularly severe. Spatial data are also required for non-infectious diseases for which environmental factors matter.

In summary, the unit of observation, which we study, is not the individual, but rather the number of individuals in zones with Covid-19 related outcomes at or during a given time period. For example, the number of people ever diagnosed with Covid-19 as of January 1, 2021, or the number of new cases diagnosed during the week ending on January 1. These are time series data that are typically updated daily or weekly, and are the spatial counterparts to national data for Covid-19 outcomes, which have featured continuously in the media since the outbreak of the pandemic.

We challenge current practice of NSA's in their response to the release of spatial Covid-19 data in three respects. First, we claim that they confuse absolute and relative confidentiality when dealing with spatial data resulting in excessive data protection where it is not mandated. Second, we challenge the response of NSA's to data protection through the practices of truncation and censoring of spatial Covid-19 data. Truncation arises when data for zones with small populations are not made public. Censoring occurs when morbidity data are grouped, e.g. morbidity during the last week is a number between 1 and 14. Whereas truncation conceals all the data, censoring reveals part of the data. We claim that in the case of Covid-19 data, truncation is applied heavy handedly, while censoring is generally unjustified. Third, we suggest that in reference to spatial Covid-19 data, NSA's have confounded individual and collective confidentiality. Recall that relative confidentiality refers to individuals, and collective confidentiality refers to statistical organizations.

The paper proceeds as follows. Section 2 addresses the unique nature of spatial data and emphasizes that personal confidentiality in such data is relative (probabilistic) and not absolute. The concepts of relative and collective confidentiality are explained in section 3, and their relationships to data truncation and censoring are elucidated. We show that while truncation may be justified under certain conditions relating to collective data protection, censoring has no obvious rationale. A review of spatial Covid-19 data availability in several countries is provided in section 4. It highlights the various data limitations applied by NSA's and underscores the very different national contexts within which data are made public. In section 5 we question the legal justification for attaching importance to relative and collective data protection. Section 6 summarizes and concludes.

Although we are concerned with universal issues in data protection policy and we provide a review of international practice, we highlight the case of Israel to illustrate our arguments. We are naturally more familiar with the intricacies of data protection where we live. However, we believe that despite some idiosyncrasies, they are not a-typical of practice elsewhere.

2. NSA's and the Nature of Spatial Covid-19 Data

2.1 Granularity and Confidentiality

NSA's traditionally conduct censuses and surveys, such as labor force surveys and income expenditure surveys, which provide detailed demographic information about individuals. They also provide geographic information. For example, in the United States the census tract block in which there are between 600 and 3,000 inhabitants is the most granular spatial zone in public use files. In the United Kingdom the statistical ward is the most granular unit and wards are merged if they have less than 1,000 inhabitants. In Israel, the statistical area is the most granular spatial unit and the populations in these zones range between 1,000 – 5.000.

NSA's seek to guarantee absolute confidentiality. Geographical or spatial data are key candidates for disclosure (Fienberg 1994, Fienberg and Willenborg 1998). Suppose, for example, that in a most granular zone occupations are recorded and there happens to be only one vet. Unless the vet's occupation is concealed it will be possible to know his or her income as well as other personal data. The public at large may not know that there is only one vet, but matters are different for other residents in the vet's zone, as well as

perhaps in neighboring zones. If there are two observationally (demographically) similar vets, each vet will know the other vet's income, and others will know their income range. Absolute confidentiality is more likely to be infringed the smaller the number of vets and the more observationally different they are. If there are 10 demographically different vets, each vet can be identified. If they are demographically identical, each vet faces a 10 percent probability of identification. NSA's anonymize the data so that such individuals cannot be identified.

There are, of course, numerous examples of data censoring motivated by absolute confidentiality This generally arises with respect to large microdata sets such as the Community Innovation Survey in the EU (Franconi and Ichim 2009) and business and household survey microdata in the US such as the BLS Current Employment Statistics or Current Population Survey (Dalton et al 2021). Another practice is top-coding where data relating to extreme values in variables such as such as income or demographic and health attributes are censored to protect the confidentiality of atypical and identifiable populations such as millionaires or the aged.

Suppose that there are a number of vets in the zone but their data include dates of birth. If there is public access to a national register of vets, which includes names and dates of birth, individual vets may be identified through triangulation. In such cases NSA's censor the data to protect their confidentiality.

These censoring practices are rightly motivated by absolute confidentiality despite the fact that the general public may have no way of revealing that there is only one vet. When the unit of observation is an aggregate such as a zip code, neighborhood or statistical area matters are different. In these zones, the probability of detection faced by individuals is 1/N where N denotes the population in the zone. The more granular the zone, the smaller is N, hence the higher is the probability of detection. There is an obvious trade-off between granularity, or spatial resolution, and relative confidentiality as measured by the probability of detection. We make two arguments. First, although not required by law, NSA's have attached importance to relative confidentiality. Second, they have set arbitrarily severe criteria regarding the trade-off between spatial resolution and relative confidentiality.

NSA's increasingly provide geocoded data of various types. These data take the form of spatial panel data in which the unit of observation has coordinates in space and time.

For example, quarterly house prices in zones (e.g. Federal Housing Finance Agency for US metropolitan statistical areas), or labor market data in zones (e.g. European Union's NUTS2 regions). These spatial zones are not too granular, so the issue of confidentiality does not arise. On the other hand, data on municipal election results, are often highly granular, as they are for some countries in the case of Covid-19.

2.2 The Nature of Spatial Units

Using zones rather than individuals as units of observation raises questions regarding the relevance of individual confidentiality in spatial data. Zones cannot be considered as 'individuals' even if their attributes such as topology and composition, are unique. Furthermore, zones vary by shape and size. These issues constitute the well-known modifiable areal unit problem (MAUP) in spatial analysis (Openshaw and Taylor 1979, Fotheringham and Wong 1991, Nelson and Brewer 2017, Tuson et 2019). MAUP highlights the arbitrary nature of spatial units and the distortions arising from the way in which space is aggregated. A related matter refers to self-selection of individuals into zones or neighborhoods (Clark 1991, Kwan 2012, Burden and Steel 2016). Individuals or firms locate in zones according to their characteristics. For example, the housing locations of individuals may reflect their physical or socio-economic amenities, including school quality, crime, parks etc. Additionally, their demographic composition changes through immigration and emigration. Hence, notions relating to the protection of individual confidentially in spatial data become obtuse.

A further issue concerning aggregating spatial zones relates to information loss. Shlomo (2010) tests the empirical impacts of aggregating or merging spatial units in an effort to preserve confidentiality. She finds that this approach generates more information loss than alternative methods for preserving confidentiality such as post-randomization probability (the PRAM mechanism) where categories of variables are changed according to a prescribed probability matrix and a stochastic selection process.

2.3 Absolute and Relative Confidentiality

Laws of confidentiality are concerned with absolute confidentiality, which involves the release of information about individuals[1] . In this event the probability of detection is one by definition. These laws do not directly address the difference between absolute confidentiality and relative, or probabilistic, confidentiality. There is an obvious qualitative difference between absolute and relative confidentiality. Sweeney (2002)

has referred to this as 'k-anonymity' in which the probability of detection is 1/k. If k equals one absolute confidentiality is at issue; if k exceeds one relative confidentiality is at issue.

As NSA's are mandated to protect the identity of individuals and as existing legislation seeks to guarantee absolute confidentiality, one might argue that by default only absolute confidentiality should be in the purview of NSA's. According to this view, if the probability of detection is one half because the number of individuals with Covid-19 is one and there are only two inhabitants in the zone, anonymity is preserved because it is impossible to determine which of the two inhabitants has Covid-19. If, instead, both inhabitants have Covid-19 it would be necessary to anonymize or de-identify the data to prevent infringement of absolute confidentiality. Whereas absolute confidentiality is uniquely defined, relative confidentiality is not.

We document below how NSA's have restricted public access to spatial Covid-19 data ostensibly on the grounds of confidentiality and data protection. For example, In Israel the Ministry of Health does not publish Covid-19 data for zones with less than 2,000 inhabitants, and if there are more than 2,000 inhabitants, it only provides uncensored data if the number of cases is at least 15. If the number is between 1 and 14 the precise number is concealed (see for example, DataGov 2021a)[2]. In statistical terms, the latter data are 'censored', whereas the former data are 'truncated'.

NSA's in other countries apply similar rules for censoring and truncation, but with different degrees of restriction. Less liberal NSA's have larger population cut-offs (3,000 instead of 2,000) and larger thresholds for the number of cases in the data (20 instead of 15). Censoring and truncation are usually justified by NSA's on the grounds of confidentiality, but they do not distinguish between absolute and relative confidentiality.

A further example of NSA's mandating excessive data protection and imposing misdirected regulation, relates to the insistence of NSA's (e.g. in Israel) on compliance with the Declaration of Helsinki. This statement outlines the ethical principles guiding medical research involving experiments with human subjects. Since Covid-19 data have not been generated experimentally, the Declaration of Helsinki is not relevant. Nevertheless, laws of individual confidentiality may be relevant if the probability of detection is large. In this case, protecting the individual's identity is not dependent on

the number of Covid-19 cases in the data because the probability of detection is 1/N for all. It is relevant, however, for collective confidentiality faced by NSAs, which obviously varies inversely with the number of cases. When confidentiality is juxtaposed with the right to freedom of information, the case for limiting public access to official statistics needs to show that the latter compromises the former. While this can be upheld for absolute confidentiality, the issue is more obscure with respect to relative confidentiality. For example, spatial data for Covid-19 are required for the epidemiological study of its spatiotemporal diffusion (Elliot et al 2020, Krisztin et al 2020, Tsori and Granek 2021), in which context more granularity is better than less. Research on the spatial diffusion of Covid-19 will inform the design of local lockdown, social distancing and 'traffic light' policy (Giannone, Paixão and Pang 2020, Narayanen et al 2020, O'Sullivan et al 2020). Also, the public has a right to know for their own protection where the incidence of Covid-19 is greater or less. Here too, more granularity is better than less.

While NSA's use of the of the Helsinki Declaration imposes an unnecessary hurdle, the directive does establish the important principle that a trade-off exists between public interest and personal privacy. Although observational data (such as spatial Covid-19 data) are not obtained through 'informed consent' including 'disclosure of personal information', nevertheless the probability of individual detection needs to be balanced against the probability of benefiting from the freedom of information. In the case of spatial data for Covid-19, the needs of science and society are very large. These include replacing national lockdown policy, for which the economic and social costs are very large, by spatial lockdown policy for which these costs are much smaller.

NSA's have enabled authorized researchers complete access to anonymous but uncensored and untruncated data in 'research rooms' using stand-alone computers and under strict supervision to prevent data leakages. More recently, 'virtual' research rooms have been developed to enable remote access to unexpurgated confidential data so that researchers do not have to be present physically (Reuter and Musuex 2010). While these simply extend the trend of increasing remote access they raise a host of issues relating to the competencies of NSA's in establishing and monitoring such facilities (EUROSTAT 2009). NSAs have also made available micro data under contract (MUC) to authorized researchers, who agree to legal stipulations and limitations. MUC files are more restricted than those available in research rooms. These welcome

developments are not germane here, where we are concerned with public use files (PUF), which are accessible to the public at large without having to undergo bureaucratic screening.

3. Concepts of Confidentiality and Techniques of Protection

In this section we define more rigorously the concepts of relative and collective confidentiality on the one hand, and truncation and censoring on the other.

3.1 Relative and Collective Confidentiality

Let θ = 1/N denote the probability of detection faced by individuals where N is the population in the zone. If the outcome applies to a subgroup of the population, e.g. adults, then N would exclude children. Let n be the number of Covid-19 outcomes (such as morbidity, hospitalizations or deaths) in the spatial zone. The probability of d detections has a binomial distribution:

$$P(d) = \binom{n}{d} \theta^d (1-\theta)^{n-d} \qquad (1)$$

The mean number of detections is nθ with variance nθ(1 - θ). Equation (1) makes the simplifying assumption that θ is the same for all subjects. After the first subject is discovered the probability of detection increases from 1/N to 1/(N - 1), and so on. Strictly speaking, therefore, P(d) has a hypergeometric distribution. However, because in the case of Covid-19 outcomes N is large relative to n, 1/(N – d) is insensitive to d. Consequently, we use equation (1) to illustrate our arguments even if it slightly under-estimates the probabilities of individual and collective detection.

Whereas individuals are naturally concerned with their risk of personal detection as expressed by θ, the statistical authorities are concerned with the probability that anyone will be detected regardless of who it might be, as expressed by 1 – P(d = 0) = P(d > 0). We refer to this probability as the "collective probability" of detection because it expresses the collective risk that at least someone will be detected. The collective probability of detection is obviously many times greater than the individual probability of detection because it varies directly with n.

If n is absolutely large, but continues to be small relative to N (as it typically does in Covid-19 data), the Poisson distribution, which is computationally simpler, provides a

good approximation to the binomial distribution, especially when n > 20 and θ < 0.05 and when n > 100 and nθ < 10. In this case equation (1) becomes:

$$P(d) = \frac{(n\theta)^d e^{-n\theta}}{d!} \qquad (2)$$

Let λ = n/N denote the incidence of Covid-19 in the population. For example, if the outcome is cumulative morbidity, λ is the proportion of the population diagnosed with Covid-19, which over three waves of Covid-19 in Israel, averaged about 0.01 or 1%. With the passage of time λ increases as new cases are diagnosed. If the outcome refers to new cases diagnosed λ = Δn/N. Notice that the mean number of detections is nθ = λ with variance λ(1-1/N). Hence, the variance varies directly with the morbidity rate and varies directly with population. As N tends to infinity, the mean equals the asymptotic variance, as expected.

Table 1 illustrates equation (1) for different values of N and n (or λ). In the first row in Table 1 there are 20 cases of Covid-19 in a population of 2,000, hence the individual probability of detection is 0.0005 or 0.05 percent and λ = 0.01 or 1 percent. The probability of collective detection faced by the statistical agency, measured by the probability of at least one detection, is 0.995 percent. (The probability of 1 detection is 0.99 percent). As expected, the probability of collective detection is many times greater than the probability of individual detection. In row 1 the probability of collective detection is 19.9 times larger than the probability of individual detection. The expected number of detections is 0.01 with standard deviation equal to 0.1. Row 2 is the same as row 1 except there are 100 cases of Covid-19 instead of 20, so λ = 0.05. The expected number of detections increases fivefold, and the probability of collective detection increases to 4.88 percent, which has increased to 97.56 times larger than the probability of individual detection.

Table 1: Individual vs Collective Risk of Detection

N	θ	n	P(d > 0)	E(d)	sd
2000	0.0005	20	0.00995	0.01	0.1
2000	0.0005	100	0.04878	0.05	0.224
4000	0.00025	40	0.009905	0.01	0.1
4000	0.00025	100	0.02469	0.025	0.158
1000	0.001	10	0.0096	0.01	0.1
1000	0.001	100	0.09521	0.1	0.316
200	0.005	2	0.0097	0.01	0.1
400	0.0025	4	0.0096	0.01	0.1
800	0.00125	8	0.0096	0.01	0.1

Based on equation (1)

In rows 3 and 4 the population is doubled to 4,000 and in rows 5 and 6 it is halved to 1,000. In the final three rows the population is less than a thousand, and the number of cases is assumed to be one percent of the population. The individual probabilities of detection vary between 0.5 percent and 0.125 percent, while the collective rates of detection are 0.97 percent.

In summary, collective rates of detection are many times larger than individual rates of detection for given rates of incidence (λ). Hence the risk of detection faced by statistical agencies, where at least one individual is detected, is much greater than the risk of detection faced by individuals. Perhaps this phenomenon motivates statistical authorities to truncate the data. If so, for given rates of incidence, Table 1 shows that the probability of collective detection is virtually independent of population size; the exposure of statistical authorities to collective detection is the same if the population is 1000 (row 5) as it is when it is 4000 (row 3). We therefore conclude that individual probabilities of detection remain small for populations less than 1000 while collective probabilities of detection are insensitive to population size.

3.2 Truncation

What would an NSA achieve if it decided to truncate the data at 2,000 instead of 1,000? For these purposes we may compare rows 1 and 5 in Table 1, which share common assumptions for $\lambda = 0.01$. First, relative confidentiality faced by individuals is much greater because the probability of personal detection is 0.1 percent when the population is 1,000 and it is 0.05 percent when the population is 2,000. However, collective confidentiality is hardly different; it is 0.96 percent when the population is 1,000 and it is 0.995 percent when the population is 2,000. Hence a more liberal NSA, which makes public data for less populated zones, decreases relative confidentiality faced by individuals, but increases collective confidentiality faced by NSAs to a much smaller extent. This difference stems from the fact that, conditional on λ, there are fewer cases of Covid-19 in less populated zones.

3.3 Censoring

In this section we now illustrate why, contrary to NSA claims, censoring is unrelated to data protection. In contrast to the foregoing, our statistical critique now draws on a real-world example. The Ministry of Health (MoH) in Israel censors the number of cases between 1 and 14. If the population is 4,000 the individual probability of detection is 0.025 percent. If the number of cases is 1, the collective probability of detection equals the individual probability of detection. If the number of cases is 14 the individual probability of detection remains unchanged, but the collective probability of detection increases to 0.0349 percent. The true probability of collective detection is bounded by these limits. Censoring makes no difference to the individual probability of detection, but why should the MoH wish to conceal the collective probability of detection?

In any event the data cease to be censored when the number of cases exceeds 14. For example, if the number of cases is 15, it becomes public knowledge that the collective probability of detection is 0.0374 percent whereas the individual probability of detection remains unchanged at 0.025 percent. So, what is the purpose of censoring the data when sooner or later the collective probability of detection is going to become public information? There is no rational reason.

Indeed, this issue is even more puzzling because MoH applies the same rules of censoring to the cumulative number of cases (n) as well as the number of new diagnoses (Δn). Initially the number of cases is zero, so the zone is 'clean'. MoH publishes this information because it believes correctly that issues of confidentiality do not arise. Suppose at some point in time t_1 Δn = 3 so n = 3. The zone ceases to be clean, but the data for n and Δn are censored because they are less than the threshold (14). At t_1 somewhere between 1- 14 cases were diagnosed. Suppose later at t_2 that Δn = 7 so that n = 10. The data continue to be censored. Nevertheless, we at least know at t_2 that n in t_1 could not have been greater than 13, therefore somewhere between 1 – 13 cases were diagnosed and in t_2 the range of n is 2 – 14. Suppose at t_3 Δn = 6 so that n = 16. The latter ceases to be censored because it exceeds 14, but the former continues to be censored. At t_3 we know that there were between 2 – 14 new diagnoses. Finally, suppose at t_4 n increases to 18 so that Δn = 2. Since the latter is less than 14 it remains censored. However, this censoring no longer matters because Δn may be calculated directly using the uncensored data for n. Despite this MoH continues to censor Δn regardless of the fact than n has ceased to be censored.

In summary, whereas truncation may, in principle, be justified in terms of relative data protection, censoring has no rationale. It creates an artificial smoke-screen, which has nothing to do with data protection either individual or collective, and which may create the impression that NSA's have something to hide. Or it may create the impression that they are irrational. Re-identification is not an issue here as zone-based Covid-19 morbidity data released by NSA's provide no other identifying characteristics of the individuals in the zone. Finally, collective confidentiality faced by NSA's varies inversely with truncation simply because there are more cases of Covid-19 in more populated zones.

4. Spatial Covid-19 Data Availability

4.1 Comparing Countries

NSA's release Covid-19 data at different levels of spatial granularity. Even within the EU there is no uniform spatial unit that serves all member states (ECDC 2021). The choice of spatial resolution has implications for confidentiality. The constraints on data availability allow us to compare across a random but representative selection of countries for which local-level data are available (Table 2). To afford comparison we standardize the different spatial units of availability to EU NUTS units. We distinguish between three types of data restrictions depending on the level of spatial resolution (Table 2).

(1) For administrative reasons there happen to be no data that are sufficiently granular for issues of confidentiality to arise. The majority of countries fall into this category for example Canada, Australia, Sweden, Germany and Italy. On the other hand, we cannot rule out that NSA's in these countries might have decided to avoid developing more granular data on the grounds of confidentiality.

(2) Granular data happen to be available, but the statistical authorities restrict their availability on the grounds of confidentiality, as in Israel, Belgium and the United Kingdom.

(3) Granular data happen to be 'incidentally' available rather than by design. This occurs in countries such as the US, France, Spain and Holland. For example, Covid-19 data are available for US counties, which typically have large populations. However, a handful of counties have zones with populations less than a thousand. Although these incidentally available data may not be useful for research into the spatial diffusion of Covid-19, they establish the principle that issues of confidentiality do not arise in small zones.

Confidentiality does not overtly arise for the first category. It is always possible, however, that it arises invisibly; the data are available to government agencies but they do not acknowledge their existence. In principle NSA's can compile such data from individual administrative records to which they have access. However, they might not have carried out this exercise, or they might not have had the necessary geocoded data

to do so. Confidentiality arises overtly for the second category. As for the third category, the statistical authorities act as if issues of confidentiality do not arise.

Spatial Covid-19 data are available for almost all countries, see for example ECDC (2021, Naqvi 2021). However, in most cases their degree of granularity is low; even the smallest spatial unit has several thousand inhabitants, if not more (Table 2). For example, in Canada the spatial units are sub-provincial area health authorities, the smallest of which have populations exceeding 10,000. In Italy the data are by province, the smallest of which (Isernia) had a population of 84,379 in 2019. In Sweden and Germany too, the data for municipalities and Landkreisen and Kreisfreien Städte are for large spatial units. The same applies to data available for 20 District Health Boards in New Zealand, and zip codes in Australia, where even in rural areas and the outback zip code populations exceed 10,000. Data are available for 47 prefectures in Japan and 154 cities and counties in South Korea, all of which have populations, which run into the 10,000s and more.

However, for some countries the spatial units for which Covid-19 data are reported have populations less than 1000. While the vast majority of US counties have populations exceeding 10,000 and at the extreme Los Angeles county has over 10m population, some counties have small populations. For example, Covid-19 data are available for Grant County in Nebraska with a population of 660 in 2018. France comprises 36,552 communes many of which have populations less than 1000 for which Covid-19 data are available. Holland comprises 355 municipalities for which Covid-19 data are available, most of which have large populations. However, some such as Schiermonnikoog have small populations (947). There are 581 Belgian municipalities of which five have populations between 1,000 – 2,000 for which Covid-19 data are available (not truncated). However, the data are censored if the number of cases is between 1-5. The same applies to Spanish municipalities, not all of which have data for Covid-19, but some municipalities such as Priego-Cuenca (pop 896) and Camaleno-Cantabria (pop 938) have small populations (Table 2).

We have already mentioned that in its public use file, the Ministry of Health in Israel truncates Covid-19 outcomes for statistical areas with populations less than 2,000, and it censors outcomes for with 1 -14 cases otherwise. The Office of National Statistics (UK) reports Covid-19 morbidity in England and Wales during the previous seven days

for 'middle layer super output areas' (MLSOA), which are sub 'lower tier local authorities'. Although MLSOAs are the most granular data available, the smallest MLSOA has 4,500 inhabitants and many have more than 10,000[3]. However, ONS suppresses these data 'in the interest of confidentiality' if the number of diagnoses is less than 3. Hence, ONS truncates data by ensuring that MLSOAs have at least 4,500 inhabitants and it censors them if morbidity is less than 3.

In summary, for the vast majority of countries spatial Covid-19 data are neither truncated nor censored because issues of confidentiality do not arise since zones have large populations. In some countries, such as Belgium, Latvia and Estonia, the data are censored or converted into ranges (see Naqvi 2021) but not truncated. In others they are truncated but not censored, and in Israel and the United Kingdom they are both censored and truncated. Finally, in countries such as the US the data are neither censored nor truncated; they are 'incidentally' unrestricted.

Table 2: Availability of Spatial Covid-19 Data by Country and Subnational Spatial Units

Country	Availability of Spatial Covid-19 Data	Spatial Unit of Availability*	NSA response
Canada	Low-level granularity	Sub-provincial area health authorities (NUTS 3)	-
Australia	Low-level granularity	Zip code (LAU)	-
New Zealand	Low-level granularity	District Health Board (NUTS 3)	-
Japan	Low-level granularity	Prefectures (NUTS 2/3)	-
S. Korea	Low-level granularity	Counties (LAU)	-
Sweden	Low-level granularity	Municipalities (LAU)	-
Germany	Low-level granularity	Landkreisen (NUTS 3)	-
Italy	Low-level granularity	Provinces (NUTS 3)	-
UK	Restricted	Middle Layer Super Output Areas (LAU)	Censoring <3 cases Truncation < 4500 pop
Belgium	Restricted	Municipalities (LAU)	Censoring <5 cases
Israel	Restricted	Statistical Areas (LAU)	Censoring <15 cases Truncation <2000 pop
US	Incidental	Counties (LAU)	Small zones unrestricted
France	Incidental	Communes (LAU)	Small zones unrestricted
Spain	Incidental	Municipalities (LAU)	Small zones unrestricted
Holland	Incidental	Municipalities (LAU)	Small zones unrestricted

*Corresponding EU NUTS spatial units in parentheses: NUTS 2 regions have roughly 0.8-3.0m inhabitants; NUTS3 regions have populations ranging from 150-800 Th); LAUs (local administrative units, previously NUTS 4 and NUTS 5 areas) have populations ranging from double digits to over 100,000 inhabitants.

4.2 Availability of Other (non-Covid19) Spatial Data

The restrictions imposed on spatial Covid-19 data do not seem to be applied to other spatial data. Election outcome data are available spatially almost universally. For

example, they are available for US counties, some of which have populations less than 1,000, as noted. In the UK they are available for all electoral wards regardless of size. The publication of election results in a ward in which as many as 90 percent voted for the Labour Party is not regarded as violating privacy, even where the electoral turn-out was very high. Election results are available for locations in Israel provided the electorate exceeds 1,000. In almost all countries, election results are available to a high degree of granularity. Although in principle there is no difference between the privacy of political preferences and individual health status, in practice statistical authorities in Great Britain, Belgium and Israel apply stricter criteria to morbidity data than they do to electoral data. On the other hand, election results are made public for reasons of democratic transparency, even where electorates are small.

In Israel the Central Bureau of Statistics (CBS) has recently started to publish data for socio-economic clusters by statistical areas. These clusters range upwards from 1 to 10 based on a variety of social and economic outcomes in these areas. However, for reasons of confidentiality, CBS truncates the data for statistical areas with less than 120 inhabitants (of which there are very few). Whereas the Ministry of Health (MoH) truncates Covid-19 data at 2,000, and the Interior Ministry truncated election results at 1,000, the Central Bureau of Statistics truncates socio-economic data at 120. Since the socioeconomic status of individuals is just as confidential as their Covid-19 status, either MoH arbitrarily attaches more importance than CBS to confidentiality, or the inconsistency results from administrative incompetence.

Another example relates to housing transactions. In Israel these require the payment of Acquisition Tax according to the price contracted. Following a successful legal challenge based on the Freedom of Information Act, the Tax Authority provides a public use file for the universe of individual house price transaction (dating back to 1989) to a very high degree of spatial granularity. Indeed, one of the purposes of this data transparency is to increase the efficiency of housing markets so that the buyers and sellers can inform themselves of recent transactions prices in neighborhoods of interest (Ben Shahar and Golan 2019). Since there are typically about 1,200 apartments in these zones, the probability of detection is much greater than it is for Covid-19 data. Moreover, the PUF contains data on housing characteristics, which increase identifiability. For houses bearing 'for sale' posters identifiability is even greater. More recently, the Tax Authority has mapped the exact locations of housing units so that it is

possible to know how much buyers paid for their housing and how much sellers received. Although individuals are not identified in these data, their neighbors know how much money they received. Similar house price data are available in Holland via Kadaster - the Dutch land registry (Kadaster 2020), in the UK from HM Land Registry (GOVUK 2020) and in the US through the Zillow's Assessor and Real Estate Database (ZTRAX 2020).

In sum, criteria for confidentiality in spatial data vary between countries for the same outcomes, and they vary within countries for different outcomes. Also, confidentiality criteria for Covid-19 outcomes vary between countries, and they vary within countries with respect to other outcomes. They even vary within countries for other medical data. For example, the Israeli Ministry of Health publishes spatial data on cancer incidence through the National Cancer Registry and censors the data in those zones with less than 50 cases annually. Considering that the rate of common cancers is about 100 per 100,000, this effectively means truncating the release of data to statistical areas with 50,000 residents. It thus seems that each statistical authority sets its own criteria. There is no coordination.

5. Stretching the Law of Data Protection

When a new phenomenon arises, such as Covid-19, providers of national statistics invent new criteria, which are supposed to protect individual confidentiality. These criteria have nothing to do with the protection of absolute confidentiality. Nor do they have much to do with relative confidentiality because in practice probabilities of individual detection are very small. At most, they may have something to do with collective confidentiality faced by NSA's because, the probability of collective detection is inevitably much greater than the probability of individual detection. Perhaps this lies behind the conservatism of NSA's in making public spatial data for Covid-19, which are sufficiently granular. By reducing the granularity of the data that they make public, NSA's directly reduce the individual probability of detection, which they believe will indirectly reduce the probability of collective detection. The comparison made above between rows 1 and 5 in Table 1 shows that this belief is false. Merging two zones with 1,000 people into one zone with 2,000 people halves the individual probability of detection but increases the relative risk of collective detection

by 3.64 percent in this numerical example. If NSA's are motivated by collective confidentiality, they should make the data more granular, not less. This means less truncation, not more.

Since the laws of confidentiality apply to persons and do not refer to collective identification, there does not seem to be a legal basis to the practice of data truncation by NSA's. The same applies *a fortiori* to the practice of data censoring by NSA's.

Laws of privacy refer to absolute confidentiality; they do not refer to relative or collective confidentiality. For example, the Law for the Protection of Privacy in Israel (1981) includes a list of offenses such as phone tapping, which clearly concern individuals. Issues of relative or collective confidentiality do not arise for phone tapping and other offenses listed. Item 9 on the list refers to the "use of information concerning individuals or its transmission to others" unless they have granted permission. This item too is concerned with absolute confidentiality. Nor has case law been concerned with infringements of relative confidentiality either in practice or in principle (Zarsky and Bar-Ziv 2019).

In 1996 the law was updated with respect to databanks containing personal data. Proprietors of databanks were required to appoint Data Security Officers (DSO's) to ensure that the law of 1981 is not infringed. Also, individuals should be given access to their own data. The law of 1996 did not introduce new concepts of confidentiality, such as relative or collective confidentiality. These concepts were introduced by the DSO's. The widespread heterogeneity to which we have drawn attention in public access to spatial Covid-19 data and other spatial data, stems from the way different DSO's interpret their mandate. It also explains how even within the same NSA different criteria are applied by different DSO's; Covid-19 data are truncated at 2,000 whereas data for cancer are truncated at approximately 50,000. In summary, legislation for data protection regarding data banks has created a vacuum filled by DSO's who have invented new concepts of confidentiality, which are interpreted arbitrarily.

As noted, NSA's have concentrated entirely on relative confidentiality and have attached less importance to the social and scientific benefits flowing from freedom of information. Historically and legally, the trade-off between freedom of information to achieve societal goals and the protection of privacy of individuals has been implemented through data de-identification or anonymization. The practical

mechanism for ensuring this privacy is invariably a variation of the classic *k*-anonymity algorithm (Sweeney 2002). This provides a framework for quantifying the likelihood of re-identification for anonymized data. A key strategy adopted by NSA's in this process is that of limited release whereby data are transformed by limiting their granularity both temporal and spatial using censoring and truncation[4]. European NSA's ascribe to the EU's General Data Protection Framework (GDPR) which offers a legal basis for issues of data privacy and data security and restricting data through limited release would seem to be consistent with that goal[5]. However, while upholding GDPR practice, data confidentiality should not be confused with the data privacy and data security mandated by the GDPR (Prewitt 2011). Data confidentiality deals with data disclosure and informed consent ("don't tell"). Data privacy addresses data collection ("don't ask") and data security deals with safeguards imposed on information that has already been collected. Confounding these issues may explain why NSA's have confounded individual and collective data confidentiality.

Different legal traditions exist with respect to protecting data confidentiality. Frameworks such as the GDRP in the EU opt for a centralized approach whereas individual states in the US set their own rules. In general confidentiality in the US is restricted to financial, genetic and medical data that are personal, whereas GDPR applies to all data including political data as well as innocuous data such as hair color. In terms of actual legislation, the traditions range from using a global approach grounded in primary legislation (Israel) to ad hoc regulation governing individual sectors such as health, communications etc as exists in the US. We agree with Zarsky and Bar-Ziv (2019) that although Israel ostensibly has a centralized, global approach to the protection of confidentiality, in practice there is extensive heterogeneity in the way the law is applied by different statistical agencies. Indeed, as we have seen, even the same statistical agency applies different criteria to different data.

When the case for data confidentiality is confronted with 'the public interest' as in the case of Covid-19, the legal tradition in Israel is rooted in individual confidentiality. Thus, Zarsky and Bar-Ziv (2019) note that anonymized personal medical data (protected under the Law of the Rights of the Patient 1996) can be released if the goal is to protect public health. Structural tension however exists in the law with respect to collective confidentiality. Here legal reading tends to an overly-constraining interpretation that results in the protection of individuals who are part of collective

entities such as geographic zones or neighborhoods. According to this interpretation, if statistical inference about individuals is based on group characteristics (the ecological fallacy issue not withstanding), then data restrictions may be justified. This group or 'attribute' disclosure (Fienberg and Willenborg 1998) arises, for example, if public data on average neighborhood earnings is expected to impact negatively on residents with earnings that are significantly different to the neighborhood average. Zarsky and Bar - Ziv (2019) note that such statistical stereotyping may challenge the laws governing individual privacy. However, laws of privacy do not stipulate that statistical stereotyping is illegal.

Further tension exists between the competing legal demands for protection of confidentiality and the societal benefits resulting from its release, such as improved medical research and enhanced quality of life. In the context of Covid-19, releasing spatial data may be construed as stigmatizing zones with high rates of contagion. This has to be juxtaposed with the need for authorities to provide accurate spatial data in order to increase trust, legitimacy and public compliance. Also, the public has the right to know where Covid-19 is particularly prevalent for their person protection. Faced with new Covid-19 data demands from cell phone tracking, geo-located purchasing and vehicle movements, some commentators see further data release without sufficient safeguards as the thin end of the wedge and a slide towards socially negative directions such as growing economic inequality and social unrest (Dwork et al 2020).

6. Conclusions

The growing demand for spatial Covid-19 data highlights some of the inconsistencies in NSA attempts to balance the competing claims for freedom of information on the one hand with protecting personal confidentiality on the other. As we show, while NSA response has varied greatly across countries, it has been consistent in confounding absolute and relative confidentiality and in failing to distinguish between individual and collective confidentiality. The result is heavy-handed NSA activity in the area of data protection. This is expressed via overly-severe data truncation and data censoring that is unrelated to data protection.

By definition, national legislation in the area of personal confidentiality relates to individual and not collective confidentiality. NSA and government ministries have

appointed DSO's with the express aim of instituting de-identification and anonymization practices to preserve personal confidentiality. With the increasing demands on NSA's to provide spatial data, DSO's have taken to filling the void unaddressed by individual confidentiality legislation, and have invented new ground-rules for collective and relative confidentiality. There is a need to regulate DSO's and to set guidelines governing their mandate.

With respect to absolute and relative confidentiality, matters are similar. In the absence of explicit legislation, which only addresses absolute anonymity (i.e by default k anonymity = 1), DSO's have again stepped into the void and determined arbitrary probabilities of detection. Whether k is 5 or 15 is not an issue that should be left to the individual discretion of DSO's. While legislation obviously cannot dictate the 'right' level for k, this is an area for which one size does not fit all. Empirical research can go a long way in providing guidelines for the formulation of consistent criteria in this field.

The above issues are pertinent to all spatial data protection whether economic, genetic or medical and not just spatial Covid-19 data. However, Covid-19 is contagious and has serious externalities and spatial spillovers, which do not apply to other diseases although they may apply to diseases subject to environmental influences such as certain forms of cancer. The public 'right to know' is particularly acute in the case of Covid-19. A freedom of information issue exists with spatial Covid-19 data that does not exist with other similar spatial data. This heightens the concern over arbitrary DSO data protection practices.

In summary, we make the following recommendations regarding the public availability of spatial Covid-19 data:

1. Data censoring should be abandoned; it serves no purpose.
2. Data truncation should be greatly curtailed. Probabilities of detection should be increased from 1 per mil to no more than 1 percent.
3. National statistical offices should regulate the ad hoc practices of DSOs.
4. Ministries of Justice should review the case for relative confidentiality.

References

Ben Shahar D and Golan R (2019) Information shock and price dispersion: A natural experiment in the housing market, *Journal of Urban Economics*,112, 70-84. https://do.orgi:10.1016/j.jue.2019.05.008

Burden S and Steel D (2016) Empirical zoning distributions for small area data, *Geographical Analysis*, 48 (4) 373–90. https://doi.org:10.1111/gean.12104

Clark, W.A.V. (1991) Residential Preferences and Neighborhood Racial Segregation: A Test of the Schelling Segregation Model, *Demography* 28, 1-19. https://doi.org/10.2307/2061333

Dalton M., Groen JA, Loewenstein MA., Piccone DS and Polivka AE (2021) The K-Shaped recovery: Examining the Diverging Fortunes of Workers in the Recovery from the Covid-19 Pandemic using Business and Household Survey Microdata, *Covid Economics*, 71, 19-58.

DataGov (2021a) *Covid-19 Data by Statistical Areas* https://data.gov.il/dataset/covid-19/resource/d07c0771-01a8-43b2-96cc-c6154e7fa9bd

DataGov (2021b) *Covid-19 Data by Sex and Age Categories* https://data.gov.il/dataset/covid-19/resource/89f61e3a-4866-4bbf-bcc1-9734e5fee58e

de Montjoye, Y.-A. et al (2018) On the privacy-conscientious use of mobile phone data. *Scientific Data*. 5:180286, https://doi.org:10.1038/sdata.2018.286

Dwork C, Karr A, Nissim K, and Vilhuber L (2020) On Privacy in the Age of COVID-19, *Journal of Privacy and Confidentiality* 10 (2). https://doi.org:10.29012/jpc.749.

Elliot RJR, Schumacher I and Withagen C (2020), Suggestions for a Covid-19 Post Pandemic Research Agenda in Environmental Economics, *Environmental and Resource Economics*, 76 (4), 1187-1213. https://doi.org/10.1007/s10640-020-00478-1

ECDC (2020) EU/EEA and UK Regional Data on Covid-19 https://www.ecdc.europa.eu/en/publications-data/sources-eueea-regional-data-covid-19

EUROSTAT (2009) *Working Session on Statistical Data Confidentiality*, Office for Official Publications of the European Communities, Luxembourg.

Fienberg SE (1994) Conflicts between the Needs of access to Statistical, Information and the Demands for Confidentiality, *Journal of Official Statistics*, 10(2), 115-132

Fienberg SE and Willenborg LCRJ (1998) Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data, *Journal of Official Statistics*, 14 (4), 337-45

Fotheringham AS and Wong DWS (1991) The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23, 1025–1044. https://doi.org/10.1068/a231025

Franconi N and Ichim D (2009) Community Innovation Survey: Comparable Dissemination, pp 11-23 in *Working Session on Statistical Data Confidentiality*, Office for Official Publications of the European Communities, Luxembourg.

Giannone E, Paixão N and Pang X (2020) The Geography of Pandemic Containment, *Covid Economics*, 52, 68-95.

GOVUK (2020) HM Land Registry: Price Paid Data https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Kadaster (2020) https://kadasterservice.nl/situaties/kadastrale-woning-gegevens

Krisztin T, Piribauer P and Wögerer M (2020) The spatial econometrics of the coronavirus pandemic, *Letters in Spatial and Resource Sciences* 13, 209–218. https://doi.org/10.1007/s12076-020-00254-1

Kwan MP (2012) The Uncertain Geographic Context Problem, *Annals of the Association of American Geographers* 102 (5), 958-968. https://doi.org/10.1080/00045608.2012.687349

Naqvi A (2021) Covid-19 European Regional Tracker, Nature: Scientific Data, 8:181, https://doi.org/10.1038/s41597-021-00950-7

Narayanan RP, Nordlund J, Pace RK, Ratnadiwakara D (2020) Demographic, jurisdictional, and spatial effects on social distancing in the United States during the COVID19 pandemic. *PLoS ONE* 15(9): https://doi.org/10.1371/journal.pone.023957

Nelson JK and Brewer CA (2017) Evaluating Data Stability in Aggregation Structures Across Spatial Scales: revisiting the Modifiable Areal Unit Problem, *Cartography and Geographic Information Science*, 44 (1), 35-50. https://doi.org/10.1080/15230406.2015.1093431

Newlands G., Lutz C., Tamo-Larrieux A, Villaronga E.F. ,Harasgama R and Scheit G (2020) Innovation under pressure: Implications for data privacy during the Covid-19 pandemic, *Big Data and Society*, https://doi.org/10.1177/2053951720976680

OECD (2020a) *Tracking and tracing COVID: Protecting privacy and data while using apps and biometrics (COVID-19)*, OECD Policy Responses to Coronavirus (Covid-19), April 2020 OECD, Paris.

OECD (2020b) *Ensuring data privacy as we battle COVID-19*, OECD Policy Responses to Coronavirus (Covid-19), April 2020, OECD, Paris

Openshaw S and Taylor PJ (1979) A million or so correlation coefficients: three experiment on the modifiable areal unit problem, pp 127-144 in Wrigley N (ed) *Statistical Applications in the Spatial Sciences*, Pion London

O'Sullivan D. Gahegan M. Exeter DJ and Adams B (2020) Spatially-explicit models for exploring COVID-19 lockdown strategies, *Transactions in GIS*, https://doi:10.1111/tgis.12660

Prewitt K (2011) Why It Matters to Distinguish Between Privacy and Confidentiality *Journal of Privacy and Confidentiality* 3(2), 41-47. https://doi.org/10.29012/jpc.v3i2.600.

Poom A.,Jarv O, Zook M and Toivonen T (2020) COVID-19 is spatial: Ensuring that mobile Big Data is used for social good, *Big Data and Society*, https://doi:10.1177/2053951720952088

Reuter W.H., and Museux JM. (2010) Establishing an Infrastructure for Remote Access to Microdata at Eurostat, in Domingo-Ferrer J., Magkos E. (eds) *Privacy in Statistical Databases*. PSD 2010. Lecture Notes in Computer Science, vol 6344. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15838-4_22

Shlomo N (2010) Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility, *Journal of Privacy and Confidentiality* 2 (1), 73-91. https://doi.org/10.29012/jpc.v2i1.584.

Spindler G and Schmechel P (2016) Personal Data and Encryption in the European General Data Protection Regulation, 7 *JIPITEC- Journal of Intellectual Property, Information Technology and E-Commerce Law*, 163, https://www.jipitec.eu/issues/jipitec-7-2-2016/4440.

Sweeney L (2002) k-Anonymity: a model for protecting privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 557-570.

Tsori Y and Granek R (2021) Epidemiological model for the inhomogeneous spatial spreading of COVID-19 and other diseases, *PLoS ONE,* doi:10.1371/journal.pone.0246056

Tuson, M., Yap, M., Kok, M.R, Murray K and Turlach B (2019) Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem, *International Journal of Health Geographics* **18,** 6 https://doi.org/10.1186/s12942-019-0170-3

Zarsky T and Bar-Ziv S (2019) Privacy's 'Identity Crisis': Regulatory Strategies in the Age of De-Identification, *Law, Society and Culture*, Vol 2, p125-166 (Hebrew) . https://ssrn.com/abstract=3350266

ZTRAX (2020) Zillow's Assessor and Real Estate Database (ZTRAX), https://www.zillow.com/research/ztrax/

Endnotes

[1] See for example the EU's General Data Protection Regulation (GDPR) Recital 26 "The principles of data protection should apply to any information concerning an identified or identifiable natural person." https://gdpr-info.eu/recitals/no-26/. Spindler and Schmechel (2016) discuss the way the GDPR addresses absolute versus relative confidentiality.

[2]This censoring is administered not only to Covid-19 data aggregated into zones but also to other aggregates such as Covid-19 data by age groups (see for example DataGov (2021b) where <15 truncation is also applied).

[3] In Scotland the spatial unit equivalent to the MLSOA is the Intermediate Zone (IZ) with a minimum population of 2,500. Public Health Scotland censors data if the number of Covid-19 cases in these zones is between 1 and 5.

[4]The limited effectiveness of these constraints on re-identification becomes ever-more pronounced in a data environment fed by geo-located mobile data. In this context recent research shows that absolute (individual) confidentiality can be compromised by a limited set of data points. For example, just 4 spatio-temporal points are enough to detect 90 percent of observations in a credit card data base of 1 million and 95 percent in a cellular phone database of 1.5 million (de Montjoye et al 2018).

[5] See GDPR, Article 89, Recitals 162-3 https://gdpr-info.eu/recitals/no-162/