Algorithmic hospital catchment area estimation using label propagation

Rob Challen

Robert Challen[1,2]; Gareth Griffith[3]; Lucas Lacasa[4]; Krasimira Tsaneva-Atanasova[1,5,6];

1) EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, Exeter, Devon, UK.
2) Somerset NHS Foundation Trust, Taunton, Somerset, UK.
3) Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK
4) School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK
5) The Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK.
6) Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.

TODO:

- Abstract
- Funding (including CHESS data set)
- Conflict of interests

Introduction

During the COVID-19 pandemic, the rapid assessment of the available capacity of a hospital and the potential demand on its services has been important in identifying geographical areas where hospital services are at risk of becoming overwhelmed. Along with epidemic dynamics, residual hospital capacity guides the imposition of public health measures such as social distancing. When assessing the load on a hospital due to COVID-19 the demand may be unevently distributed in space and rapidly changing in time. Available capacity may be influenced by multiple factors, including staff availability. At the same time there may be fundamental changes to health provision in the acute response of the pandemic, with for example the cancellation of routine operations. In the early epidemic in the UK, for example, there was block booking of private health care providers to assist the NHS [1], and the rapid creation of large scale field hospitals [2]. In previous work we examined the potential for redirecting patients from one region to another to balance the load of health care provision [3] and we have observed this phenomenon as intensive care units reach capacity [4]. When we consider both the change in provision of services and the redistribution of patients, there is a potential need to redefine the demographic and geographic profiles of health care service providers ("catchment areas" and "catchment populations") [5] to allow for effective planning.

The catchment area or population of a hospital is a broad concept which serves a number of purposes, such as:

- Definition of the primary population of a hospital (and their demographics) for strategic planning purposes [6].
- Definition of higher level organizational structures and collaborative networks [7].
- Identification of areas with under, or over provision of services

- Calculation (and visualisation) of incidence and prevalence of disease from hospital reported statistics (identifying the denominator) [8] and hence admission rates per head of population.
- Preferred routing of patients to hospitals for optimizing specific services.

There are two general approaches to modelling catchment areas which we will discuss in detail - activity based or algorithmic approaches. Algorithmic approaches are based solely on population level information about regional demographics and hospital capacity. Activity based approaches minimally require data on hospital activity across all the region at an individual level, such as individual patient admission records.

Either of these individual modelling approaches result in a hospital catchment area that is either overlapping or non-overlapping. An overlapping output may reflect the fact that patients may have a choice in the use of the services, and that a range of individually varying predictors influence individuals' capacity and willingness to adhere to arbitrarily imposed boundaries. It may also reflect a fundamental organization of the service, for example the networks of critical care [4], in which some activity of a hospital caters directly for the local population, but other activity is conducted supporting other regional hospitals. As such overlapping approaches may better reflect reality, but non-overlapping outputs are often a necessary simplification for secondary analyses, where cross-classification is not specifiable [9]. It is often desirable for secondary analysis that boundaries align with geographical and organizational boundaries, but non-overlapping outputs are more at risk of mis-classification, and this tends to be spatially uneven, clustering at the fringes of the imposed boundaries [10].

The simplest algorithmic approaches involve straight line distance weighted to a measure of the size of hospital [11]. This can be extended by models which use an analogy to gravity to calculate the potential field of every hospital, based on both capacity (e.g. beds) and demand (e.g. patients) [11–13]. The resulting potentials may be cut off at a specified value, or where they are exceeded by another hospitals potential, to produce either overlapping or non-overlapping outputs. Such algorithmic approaches may not respect geographical or existing organizational boundaries, but they can be used to model hypothetical scenarios, such as the impact of creating a new hospital. Further details of the range of different models that have been proposed have been previously published [5,8].

Activity based models began with the proportional flow, or Norris-Bailey, model [14,15] which examines the proportion of patients from an area visiting a particular hospital versus the proportion of patients in an area who visit any health care provider. An extension of this was recently used to define catchment areas for major injury following acute trauma [16]. More recently modern statistical approaches have been applied to the same basic activity data including k-Means classification [8], Bayesian regression modelling. [6] or Markov Multiscale Community Detection [7,17]. Whilst arguably providing a more accurate reflection of reality, activity based models are predicated on the availability and recency of activity data, which may exhibit historical or cultural biases. Depending on the purpose of the catchment area such historical bias may or may not be desirable [8].

Estimation of hospital catchment areas is a simplification of a complex logistical and organizational problem. In England, for example, hospital sites are typically grouped into single organizational units (NHS trusts) which report combined activity. Thus a single unit of health-care provision (NHS trust) may have a range of physical locations, not all of which offer the full range of services. ICU provision is often focused in a single hospital in an NHS Trust, whereas acute or step-down beds may be distributed across multiple sites. Some specialist services, such as intensive care, also may be unevenly distributed, and larger units used as "tertiary referral centres" which take in more complex patients from a wider geographical area.

In the early phase of the COVID-19 pandemic, a rapid estimate was needed of the potential demand on intensive care services as a result of observed and forecast infections, in the context of a changing landscape of health service provision. At this point, there was no comparable data with which to drive activity based models, and volatile estimates of hospital capacity. In order to plan provision of additional ventilators and high dependency beds, we needed a model of geographical catchment areas that could be used to translate regional epidemiological models of infections into a prediction of future admissions to individual hospitals, taking into account the regional demographics, and an estimate of the expected level of care the patients would need. Such a catchment area model must interface with existing spatial boundaries implemented in epidemiological models and publicly available demographic estimates, and fulfil the following criteria:

- Allow a clean one way mapping from fine grained geographic regions (e.g. from regional demographic estimates or epidemiological models) to the coarse grained administrative hospital region.
- Provide contiguous and realistic subdivisions of geographies relating to a single hospital or to a hospital group.
- Provide areas that are determined by the capacity of hospital at different levels of care provision, and the density of local population, or anticipated size of outbreak in the local population.
- Create regions of approximately equal local supply (e.g. beds) and demand (e.g. patients) at boundaries.
- Respect crude physical geographical boundaries, such as large rivers.
- Flexible in that it can be recomputed rapidly if the background parameters change, for example, a regional outbreak or provision of additional hospitals, in a way that is not dependant on individual level activity data.

In this work we present the solution we developed for this problem, and introduce a novel algorithmic catchment area model which is specifically designed to meet the needs of the COVID-19 pandemic as described above. This model is inspired by label propagation techniques used for community detection in networks [18,19]. The paper is presented as follows; firstly we introduce the algorithm, secondly we describe some illustrative examples, and thirdly we qualitatively compare the output of the algorithm to both manually created organizational boundaries, and to observed patient ICU admissions during the first wave of the COVID 19 pandemic.

Methods and materials

This section consists of 3 parts: a detailed description of the algorithmic catchment area model, a description of the data used to create initial outputs from the model, and a description of initial assessment of the model against available data.

Algorithm

The algorithm is inspired by label propagation network clustering, where labels correspond to the supply of a service, and the nodes in the network correspond to the demand for the service. For illustrative purposes in this paper we will focus on the example of hospitals, where the "supply" is provision of hospital beds, the "demand" is the population size, and the "network" is the neighbourhood of geographical areas under consideration.

To connect supply and demand, or hospital beds to population size, the algorithm propagates a number of labels, each representing the source of supply (e.g. the hospital), through the geographical network, at a rate defined both both the size of the supply (e.g. beds in each hospital), and the demand for the service (e.g. the population) within the areas the label has already propagating to. Thus as demand outstrips supply from a particular source the rate of label propagation associated with that source decreases.

We assume the whole geographical region under consideration can be represented as a mathematical graph, G and is divided into N smaller regions, represented by the vertices V (where $V = V_n, n = 1, 2, \ldots, N$) each with known population of size $D(V_n)$.

We define M hospitals located at the geographical points P (where $P = P_m, m = 1, 2, 3 \ldots M$), and with capacity to supply $S(P_m)$ beds. Typically there are fewer hospitals than regions ($M << N$). We constrain P_m such that no more than one P_m is found within any given V, i.e. each small region hosts no more than one hospital. In practice the assumption that a maximum of one hospital is found in each region is occasionally not true. When this does happen, we preprocess the data to combine hospitals that are located together into a single entity.

The connections of neighbouring regions of any area V_x are defined by $E_x = \nu(V_x)$, and likewise the set of neighbouring vertices of any subgraph G_y are defined by $E_y = \nu(G_y)$. These quantities are readily calculated

using the geographical intersection of different areas and various algorithms exist to calculate these from geo-spatial data [20,21].

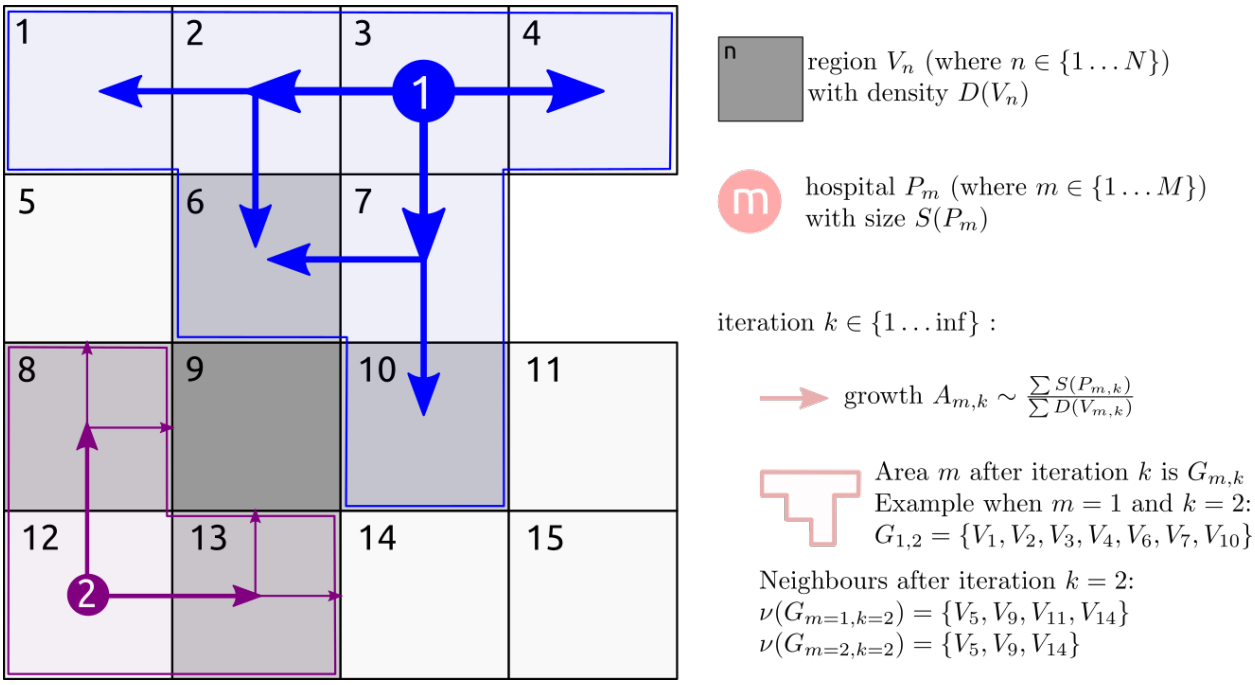

Figure 1: The principles of our proposed label propagation algorithm. The association of a hospital with a region propagates from the hospital location (P) into the different regions (V) at a rate depending on the hospital capacity S(P) and the population of the region, D(V), at each round of the iteration (k). The direction of spread is determined by the geographical neighbourhood of each region V

Our goal is to divide the graph G into M labelled sub-graphs G_m such that the sub-graphs are connected, and that neighbouring sub-graphs have similar bed availability per unit population ($\frac{\sum S_m}{\sum D_m}$). We do this by assigning a score for each combination of region and hospital, which is initially zero. For every iteration of the algorithm this score is incremented in any unlabelled region that neighbours a region that has been labelled (i.e. assigned to a specific hospital). The score is increased by a small amount determined by the ratio of supply (hospital beds) available, and demand (population to be served) in the regions assigned to that hospital. Thus labels propagate more quickly from points with a high capacity, through regions with a low population density than vice-versa. The first label to propagate to a given area, and for which the score is above a threshold is defined as the "supplier" for that area, which is labelled as such. This enforces that

each region is served by only one hospital.

Algorithm 1: A weighted label propagation algorithm for matching geographical supply to demand

Input : V_N - the N regions of demand as a set of geographical polygons
Input : $D(V_n)$ - the density of demand in any given region as a function of the region V_n
Input : P_M - a set of M labelled suppliers as a set of geographical points
Input : $S(P_m)$ - the capacity of supply at any given supply point as a function of the supplier P_m
Input : C_{growth} - a rate constant defining rate of label propagation
Output: G_M - M labelled subgraphs of graph G, relating to the catchment areas of suppliers P_M

- define G as the graph consisting of geographical regions V_N, connected by edges, E_N, given by their geographical neighbours $\nu(V_N)$:
$E_N \leftarrow \nu(V_N)$;
$G \leftarrow (V_N, E_N)$;
- define V_M and $V_{M,0}^{new}$ as the geographic regions of G serviced by points P_M, and $G_{M,0}$ as a set of labelled sub-graphs (also initially consisting solely of the vertices V_M):
$V_M \leftarrow G \cap P_M$;
$V_{M,0}^{new} \leftarrow V_M$; $G_{M,0} \leftarrow V_M$;
- define the initial unlabelled set of vertices:
$U_0 \leftarrow \neg V_M$;
- define the initial un-labelled neighbours of labelled sub-graphs, G_M:
$U_{M,0} \leftarrow \nu(V_M)$;
- define an accumulated growth score for each un-labelled neighbour $U_{M,0}$ of each $G_{M,0}$:
$A_{U_{M,0}} \leftarrow 0$;
$k \leftarrow 0$;
- execute the loop while there are still unlabelled vertices and there exist some unlabelled neighbours of labelled vertices
while $|U_k| > 0$ *and* $|U_{M,k}| > 0$ **do**
  $k \leftarrow k + 1$;
  - define the un-labelled vertices as the set of V not contained in any of $G_{M,k-1}$:
  $U_k \leftarrow \neg G_{M,k-1}$;
  - define the un-labelled neighbours of $G_{M,k-1}$ as $U_{M,k}$ as the previously unlabelled neighbours and the neighbours of the most recently labelled neighbours $V_{M,k-1}^{new}$:
  $U_{M,k} \leftarrow U_{M,k-1} \cup (U_k \cap \nu(V_{M,k-1}^{new}))$;
  - define the reserve capacity, R_M, to supply existing labelled, $G_{M,k-1}$, and un-labelled neighbours $U_{M,k}$, as:
  $R_M \leftarrow \frac{S(P_M)}{D(U_{M,k} \cup G_{M,k-1})}$;
  - for unlabelled areas only, update the accumulated growth score, $A_{U_{M,k}}$, with the normalised rank of the reserve capacity and multiplied by a constant $C_{growth} > 1$ representing the speed at which the accumulated growth score increases in all areas:
  $R_{M,k} \leftarrow R_m\{m \in U_{M,k}\}$;
  $A_{U_{M,k}} \leftarrow A_{U_{M,k-1}} + C_{growth} \times \text{rank}(R_{M,k})/|R_{M,k}|$;
  - for all the un-labelled vertices, select the label M, with the highest score, and if the accumulated score has reached the threshold of 1, incorporate it into the labelled sub-graph, $G_{M,k-1}$:
  $A_{U_k}^{max} = \max(A_{U_{m,k}}, m \in M)$;
  $V_{M,k}^{new} \leftarrow U_{M,k} \in \{A_{U_k}^{max} > 1\}$;
  $G_{M,k} \leftarrow G_{M,k-1} \cup V_{M,k}^{new}$;
  $U_{M,k+1} \leftarrow U_{M,k} \cap \neg V_{M,k}^{new}$;
end
return $G_{M,k}$

Qualitative testing data

The algorithm requires an estimate of demand, in this case we used population counts, a geographical network and an estimate of supply, in this case hospital capacity data.

For the UK there are detailed estimates of the population at granular geographic detail (lower super output area - LSOA11) available from the Office of National Statistics (ONS) for England and Wales, and the National Records Service (NRS) in Scotland [22,23]. These population estimates are available by single year of age for each area. These are combined to create a single figure for the adult population of each small geographic area.

Each geographical area is associated with a boundary file also provided by the ONS and NRS [24,25].

To estimate the capacity of hospitals we used a range of primary sources (described in the supplementary materials) to manually compile a list of NHS and independent hospital sites. When not provided in the primary sources, we identified their geographical locations from their postcode, and we estimated bed numbers from both a combination of published NHS statistics and from daily COVID-19 situation reports from early April 2020, provided by the NHS. The situation reports detailed both available beds at this point in time but also gave an indication of maximum surge capacity for high dependency beds. These data were manually curated and are indicative of the state of the NHS at maximal readiness. Bed state estimates for independent hospital providers were also available through the situation reports.

In Northern Ireland, population estimates were not available at a similar geographical resolution as the ONS and NRS sources, and we are unaware of any publicly available hospital capacity estimates. They were therefore not included in this analysis.

Validation data

There is no ground truth for the catchment areas for hospitals in the NHS during the COVID-19 pandemic, so we have not been able to rigorously validate the accuracy of our method. We did however compare the output of the algorithm two data sets that become available to give us a qualitative indication of the performance of the algorithm.

The first is a list of postcodes associated with administratively defined catchment areas for 4 NHS trusts in the South West which was provided to us by the NHS. The postcode areas were converted to LSOA codes using mapping files provided by the ONS (Source: Office for National Statistics licensed under the Open Government Licence v.3.0, containing OS data © Crown copyright and database right 2020 and Royal Mail data © Royal Mail copyright and database right 2020) [26]. The administratively defined catchment areas were taken as "observations" of the association between LSOA codes and hospital catchments. We compared this to the "predictions" of the LSOA associated with each of the 4 hospitals derived from the algorithm.

Secondly, at a much later stage of the pandemic, we were able to identify the coarse location (partial UK postcode, also known as outcode) from a list of admitted intensive care patients provided as part of the CHESS dataset [27]. We used an outcode map [28], LSOA demographic estimates, and an areal interpolation [29] to generate an estimate of demographics for each outcode. Using this outcode based regional population estimate, outcode boundary shapes, and the manually curated high dependency unit capacity estimates we can calculate an outcode based catchment area estimate.

We use outcodes of admitted ICU patients, and the data on the NHS trust they were admitted to as "observations" and, using the outcode based catchment area estimate, we compare to a "prediction" of the expected NHS trust for the admitted patients given the patients outcode location.

The detail of the original data sources we used in presented in the supplementary material, not all of which are publicly available. The algorithm is implemented as an R package [30] and this contains both the manually curated hospital capacity and data pertaining derived demographics data described here.

Results

Qualitative testing results

The results presented in this section qualitatively test the algorithm to determine whether it is producing catchment area regions that are geographically contiguous, aligned with existing demographic boundaries, respect coarse geographical boundaries such as large rivers. The catchment areas should also produce estimates that minimise differences in the level of service provision from area to area, and we expect the overall regional variation of supply versus demand to be locally smooth. Figure 2 shows a catchment area based on individual hospitals that offered high dependency beds during April 2020, and a regional demand based on population estimates of adults in lower super output areas. The resulting set of catchment areas presented in panel A and C behave as desired in terms of the geographical properties. They also produce a fairly uniform density of high dependency bed provision per capita population, from region to region, as seen in panel B, however in areas where there are high densities of hospitals such as London where the algorithm, by design, cannot propagate from centrally located hospitals past more peripheral hospitals, leading to small numbers of areas with high provision per head of population. This is discussed further below.

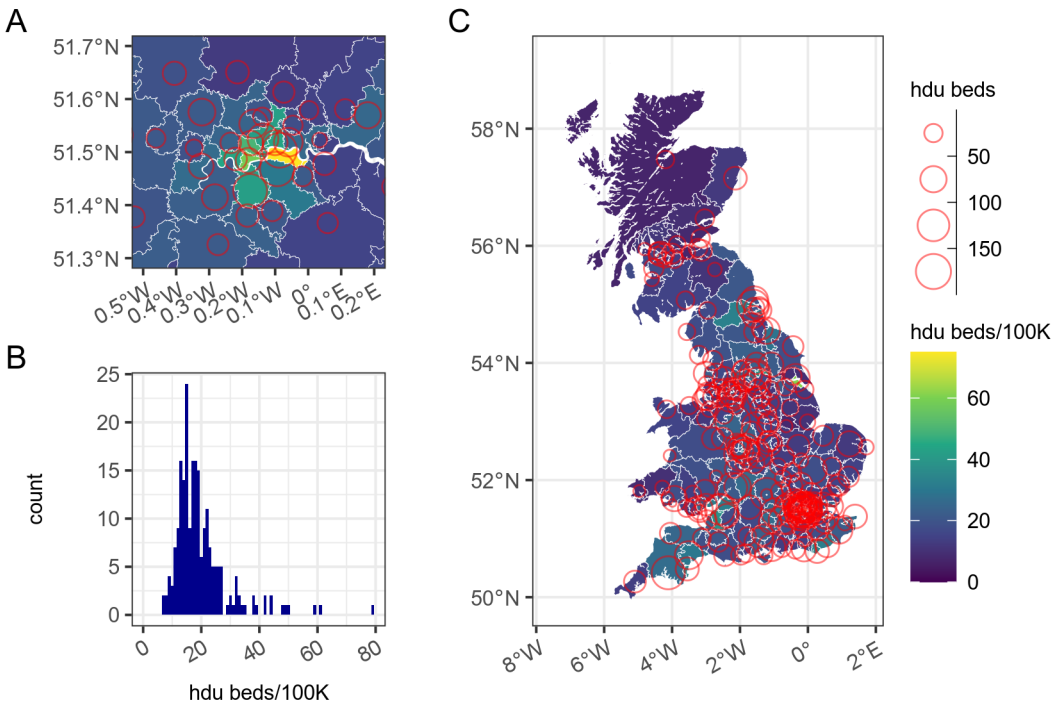

Figure 2: Panels A and C show a LSOA based catchment area map estimated from the high dependency bed state in the UK in early April 2020, with catchment area boundaries shown in white. Red circles are NHS hospital sites with high dependency capacity. Map source: Office for National Statistics licensed under the Open Government Licence v.3.0, Contains OS data © Crown copyright and database right 2020. Panel B shows the distribution of high dependency beds per 100K population for each of the catchment areas defined by the algorithm.

Further qualitative investigation of the properties of the algorithm are shown in Figure 3 where we see more regional detail of the same algorithm applied this time to general hospital beds rather than high dependency beds. Panel A shows the boundaries of the estimated catchment areas in white against the population density of a small area of the South West of England containing three hospitals (Plymouth, Torbay and the Royal Devon and Exeter hospitals). We can see in this example the extent of the catchment area to the South of Torbay is defined by the Dart river estuary, thus respecting such geographical boundaries.

Figure 3 panel B shows details about the progression of the algorithm from one iteration to the next, as labels propagate from each of the hospitals into the surrounding areas until encountering another catchment area. As we expect from the design the algorithm is seen to spread from hospital sites quickly through areas of low population density (panel A), such as the countryside surrounding Plymouth in the bottom left, and more slowly through areas of higher population density such as the areas surrounding Torbay in the middle right.

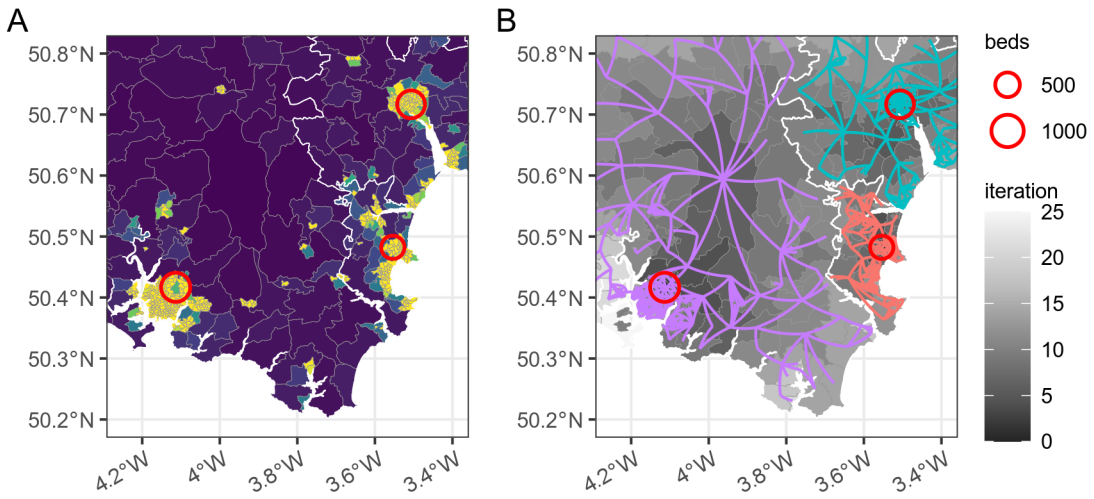

Figure 3: Detail LSOA based catchment area map for NHS trusts estimated from the general hospital bed states in the UK in early April 2020. Red circles are NHS hospital sites. In panel A the fill represents a relative measure of regional population density, with yellow areas being high density in and around cities. In Panel B the same areas are shown but this time the fill shows the iteration number at which the algorithm labelled a specific area, and the propagation of the algorithm by arrows. Map source: Office for National Statistics licensed under the Open Government Licence v.3.0, Contains OS data © Crown copyright and database right 2020

Validation

In this section we look quantitatively and qualitatively at the comparison between existing administratively defined catchment areas and the output of our algorithm for general hospital beds. The administrative assignments of postcode areas to one four NHS trusts, the Royal Devon and Exeter (RD&E), Torbay, Plymouth (Plymth) and North Devon NHS trust (N Devn) are shown in Figure 4 panel A. In comparison the estimated catchment areas resulting from the the label propagation algorithm are shown in Figure 4 panel B. There are differences in the geographical areas covered between the administratively assigned and the label propagation catchment areas, with, for example, the catchment area of Plymouth hospitals being much larger in our estimates than the administrative areas would suggest. Looking at Figure 3 panel A, however reminds us that the areas where there are major differences have low population density, and the geographical subdivisions in these regions, which have a relationship to population density, have a correspondingly large area. Therefore small differences in terms of population served may be associated with a large difference in terms of geographical area. Figure 4 panel C shows the number of areas for which there is agreement and disagreement between the two methods of assigning an area to a hospital, and the differences between the

2 methods are less striking when we consider the numbers of areas (panel C) rather than the geographical areas (panel A and B).

The use of the administratively defined regions is somewhat flawed as they do not represent a ground truth for the catchment areas of the hospitals, and as such a quantitative analysis is limited to an evaluation of the overall agreement between the two estimates which was 85.4% for the matches between the 4 hospitals, or 85.4% when including a fifth "other" category including the large number of unassigned areas over the whole of the rest of the country.

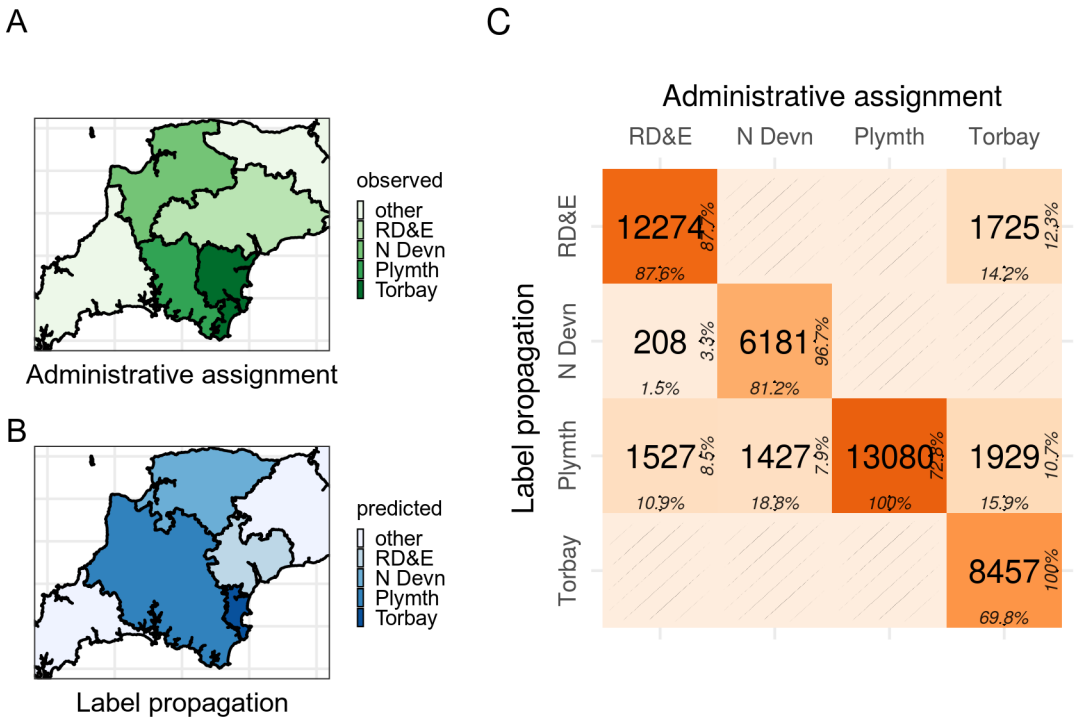

Figure 4: Comparison of an administrative catchment area (panel A) with one estimated by label propagation (panel B). A contingency table for the comparison of the 4 regions included in the administrative assignment is in panel C

Our last piece of evaluation used the observed admissions from intensive care units for which we have a coarse grained patient address in terms of the outcode, and the hospital to which the patient was admitted, this is essentially a patient level activity measure as we described in the introduction. We compare these observations to predictions made by the label propagation algorithm, for the expected hospital associated with a patient from any given outcode. As there are 215 under consideration this forms a large contingency matrix which is not shown here. In summary however we find across the whole country exact agreement between the observed location of hospital admission and the predicted location of hospital admission based on the label propagation catchment area in 12188 out of 17141 cases. This give us an accuracy of 71.1%, and the Matthew's correlation coefficient was 0.71. This lower accuracy is partly the result of the many possible output classes for predictions (one for each NHS trust).

Table 1: The top 10 misclassified hospitals

Trust	Errors
St George's University Hospitals Nhs Foundation Trust	276
Manchester University Nhs Foundation Trust	274
Cambridge University Hospitals Nhs Foundation Trust	233
London North West University Healthcare Nhs Trust	232
The Newcastle Upon Tyne Hospitals Nhs Foundation Trust	178
Guy's And St Thomas' Nhs Foundation Trust	167
King's College Hospital Nhs Foundation Trust	166
Walsall Healthcare Nhs Trust	138
Imperial College Healthcare Nhs Trust	122
Kingston Hospital Nhs Foundation Trust	111

In Table 1 we qualitatively examine the top ten hospitals that the ITU patients are actually admitted to, when the label propagation algorithm expected them to be admitted elsewhere, and mis-classified them. These represent 1897 (38.3%) of the mis-classifications overall. These 10 hospitals are all major tertiary referral intensive care units, or specialist centres. This result is consistent with both the possibilities that severely ill patients may end up in specialist centres rather than their closest hospital for treatment, or that in the event of a large surge in cases, patients may overflow from smaller to larger intensive care units. Both of these could result in the kind of mis-classification of these patients by the label propagation algorithm, as we see here.

Discussion

We have presented an algorithm for rapidly estimating hospital catchment areas for use when activity data is not available. We demonstrate how the output responds to the different capacities of the different levels of care provided (e.g. high dependency versus general hospital beds). We present catchment areas calculated using population size as demand, and total hospital beds as supply. This algorithm may be useful for longer term strategic planning, but was conceptualized as part of an acute response to COVID-19 outbreak. In this case we can use the different parameters for demand, for example local COVID-19 infection prevalence, and different parameters for supply, for example availability to staffed hospital beds. Our approach is novel in that it allows adaptation of local service provision to predictions of disease prevalence from epidemiological models of COVID-19 and real time bed states provided by NHS trusts. This allows us to model the degree of elasticity in the system to absorb localised shocks, caused by regional outbreaks, it helps us to develop a better concept of when services are being at risk of becoming overwhelmed, and allow routing of new admissions away from overloaded hospitals.

Benchmarking our algorithm against administrative boundaries produced a 85.4% agreement. Further quantitative assessment of the performance of our algorithm is limited by the lack of a ground truth. The finding that naive application of our algorithm to real world patient admission classifies only 71.1% correctly is explained by two things, firstly accuracy at boundaries decreases as the number of boundaries increases [10] and secondly the fact that all of the top 10 mis-classified trusts are major tertiary referral centres, suggesting that our constraint that catchment areas should be non overlapping is not borne out in reality for this cases. In our comparison against administrative areas, we find that some areas may be mis-classified where our algorithm propagates quickly through less densely populated regions. This is a limitation as those regions are also likely to be poorly served by transport links, and we do not consider travel time in our approach. Adding a travel time penalty to the rate of label spread into the model is possible given some estimate of the ease of transport within and between regions, and this is an area of future work.

Overlapping catchment areas could be modeled by multiple layers of non-overlapping catchment areas. When we consider the provision of intensive care services in the UK during COVID-19, we propose there are at least 3 layers of hospital service provision: there is a local service, which provides care for patients from nearby. A subset of hospitals additionally provide a regional, or tertiary referral, service layer which takes sicker patients from neighbouring hospitals in larger areas. The final layer is a crisis overflow layer provided

by the NHS Nightingale field hospitals [2]. Each of these layers may be considered to have somewhat independent catchment areas. We propose that dividing the larger hospitals into local and regional services and considering the tertiary referral network as a second layer, with its own larger catchment area would improve the performance of the algorithm against real activity data. In such a layered model of catchment areas there is interplay between local layer demand for hospital beds and capacity for regional tertiary care provision, which will dynamically affect the "catchment area" for regional tertiary care provision, potentially on a day to day basis. In previous work we looked at the opportunities for balancing the load between different hospitals [31] when transferring COVID-19 patients away from overloaded areas, however moving unwell patients between hospitals is ideally minimized. With this algorithm we enable the dynamic re-specification of local service catchment areas and hospital tertiary referral networks, based on evolving demand. Coupled with flexible load sharing has interesting potential to model or influence patient admissions around the whole hospital network.

Hospital capacity is difficult to accurately estimate. During this work we encountered many of the uncertainties that influence capacity. The ability of a hospital to provide a bed to a patient depends on a multitude of factors, including staff availability, which may vary during the different stages of the pandemic. The ability of hospitals to absorb large numbers of emergency patients by reconfiguring their service provision (e.g cancelling routine operations) and providing overflow or "surge" high dependency capacity for short periods of time makes putting a single number on hospital capacity difficult. The ability to recalculate catchment areas based on changing assumptions around capacity is a strength of our approach, and in the future could be used to analyse the impact of instroducing new capacity into the hospital system.

There are opportunities to extend our algorithm. The general approach of label propagation in networks has been more widely studied and newer approaches described [18,32,33] which allow overlapping communities. This may address some of the issues described above. These are appealing and a possible avenue for future extension of the algorithm. There persists however an open question about whether the overlapping nature of hospital service provision observed in activity data is not really a reflection of patient choice, but actually the result of subtly different services, or different levels of service, being provided by different hospitals to different catchment areas. Thus a specialist cancer hospital close to a specialist paediatric hospital will have geographically overlapping catchment areas, but in reality these hospitals are not providing the same service to the same population. This line of argument suggests that the concept of a single overlapping hospital catchment area is also an oversimplificiation, and when we take into account the hetrogeneity of different services offered by a hospital, we propose that a hospital's overall catchment area may be well modeled by a collection of non-overlapping catchment layers.

Conclusions

This label propagation algorithm for estimating hospital catchment areas is a pragmatic solution to determining geographical and demographic subsets of the population when there is no previous activity data available. It suits situations where the level of service provision and demand on the hospital system is dynamic, as has been the case in the COVID-19 pandemic. The algorithm is simple and satisfies the major criteria we set out in the introduction, in that it provides a mapping from low level geographic regions which provide contiguous and realistic subdivisions of geographies relating to a single hospital or to a group of hospitals. The areas are determined by the capacity of the hospital and the density of local population, and are approximately equal in terms of local supply (e.g. beds) and demand (e.g. patients) at boundaries.

The algorithm depends solely on data reflecting supply and geographical demand for a service, and as such is quite generic and potentially more widely applicable outside of medicine. Although we have discussed catchment areas in terms of the capacity of hospital beds, and demand of local populations, there is nothing to prevent us defining capacity in any other way - a heuristic on staffing levels may be appropriate, or in different contexts, availability of medical imaging devices. Likewise, demand may be refined to reflect sub-populations at risk of disease, or may even be the output of a predictive model. As such our approach is applicable to a wide variety of problems.

References

1. 2020 Coronavirus: Thousands of extra hospital beds and staff. *BBC News: UK*, 21 March. See https://www.bbc.com/news/uk-51989183.

2. 2020 Coronavirus: Nightingale Hospital opens at London's ExCel centre. *BBC News: UK*, 3 April. See https://www.bbc.com/news/uk-52150598.

3. Lacasa L, Challen R, Brooks-Pollock E, Danon L. 2020 A flexible method for optimising sharing of healthcare resources and demand in the context of the COVID-19 pandemic. *PLOS ONE* **15**, e0241027. (doi:10.1371/journal.pone.0241027)

4. Pett E *et al.* 2020 Critical care transfers and COVID-19: Managing capacity challenges through critical care networks. *Journal of the Intensive Care Society*, 1751143720980270. (doi:10.1177/1751143720980270)

5. Jones S, Wardlaw J, Crouch S, Carolan M. 2011 Modelling catchment areas for secondary care providers: A case study. *Health Care Manag Sci* **14**, 253–261. (doi:10.1007/s10729-011-9154-y)

6. Wang A, Wheeler DC. 2015 Catchment Area Analysis Using Bayesian Regression Modeling. *Cancer Inform* **14s2**, CIN.S17297. (doi:10.4137/CIN.S17297)

7. Clarke JM, Barahona M, Darzi AW. 2019 Defining Hospital Catchment Areas Using Multiscale Community Detection: A Case Study for Planned Orthopaedic Care in England. *bioRxiv*, 619692. (doi:10.1101/619692)

8. Gilmour SJ. 2010 Identification of Hospital Catchment Areas Using Clustering: An Example from the NHS. *Health Services Research* **45**, 497–513. (doi:10.1111/j.1475-6773.2009.01069.x)

9. Jones K, Johnston R, Manley D, Owen D, Charlton C. 2015 Ethnic Residential Segregation: A Multilevel, Multigroup, Multiscale Approach Exemplified by London in 2011. *Demography* **52**, 1995–2019. (doi:10.1007/s13524-015-0430-1)

10. Arcaya M, Brewster M, Zigler CM, Subramanian SV. 2012 Area variations in health: A spatial multilevel modeling approach. *Health Place* **18**, 824–831. (doi:10.1016/j.healthplace.2012.03.010)

11. Reilly WJ. 1931 *The law of retail gravitation,* New York: W.J. Reilly.

12. Huff DL. 1964 Defining and Estimating a Trading Area. *Journal of Marketing* **28**, 34–38. (doi:10.1177/002224296402800307)

13. Stewart JQ. 1941 An Inverse Distance Variation for Certain Social Influences. *Science* **93**, 89–90. (doi:10.1126/science.93.2404.89)

14. Bailey NTJ. 1956 Statistics in Hospital Planning and Design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **5**, 146–157. (doi:10.2307/2985416)

15. Norris V. 1952 Role of Statistics in Regional Hospital Planning. *Br Med J* **1**, 129–133.

16. Alexandrescu R, O'Brien SJ, Lyons RA, Lecky FE, Trauma Audit, esearch Network. 2008 A proposed approach in defining population-based rates of major injury from a trauma registry dataset: Delineation of hospital catchment areas (I). *BMC Health Serv Res* **8**, 80. (doi:10.1186/1472-6963-8-80)

17. Clarke J, Beaney T, Majeed A, Darzi A, Barahona M. 2020 Identifying naturally occurring communities of primary care providers in the English National Health Service in London. *BMJ Open* **10**, e036504. (doi:10.1136/bmjopen-2019-036504)

18. Xie J, Szymanski BK. 2011 Community detection using a neighborhood strength driven Label Propagation Algorithm. In *2011 IEEE Network Science Workshop*, pp. 188–195. (doi:10.1109/NSW.2011.6004645)

19. Xie J, Szymanski BK. 2013 LabelRank: A stabilized label propagation algorithm for community detection in networks. In *2013 IEEE 2nd Network Science Workshop (NSW)*, pp. 138–143. (doi:10.1109/NSW.2013.6609210)

20. Bivand R, Rundel C, Pebesma E, Stuetz R, Hufthammer KO, Giraudoux P, Davis M, Santilli S. 2020 *Rgeos: Interface to Geometry Engine - Open Source ('GEOS')*. See https://CRAN.R-project.org/package=rgeos.

21. Pebesma E. 2018 Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**, 439–446.

22. In press. Population estimates - Office for National Statistics. See https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates (accessed on 17 November 2020).

23. Scotland Web Team NR of. 2013 National Records of Scotland. See /statistics-and-data/statistics/statistics-by-theme/population/population-estimates/2011-based-special-area-population-estimates/small-area-population-estimates/time-series (accessed on 7 December 2020).

24. In press. Open Geography Portal. See https://geoportal.statistics.gov.uk/ (accessed on 17 November 2020).

25. SpatialData.gov.scot SG. 2020 Data Zone Boundaries 2011. See https://data.gov.uk/dataset/ab9f1f20-3b7f-4efa-9bd2-239acf63b540/data-zone-boundaries-2011 (accessed on 7 December 2020).

26. In press. Open Geography portal. See https://geoportal.statistics.gov.uk/ (accessed on 18 December 2020).

27. In press. SGSS and CHESS data. See https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/data-provision-notices-dpns/sgss-and-chess-data (accessed on 18 December 2020).

28. In press. Open Door Logistics - Intelligent software for vehicle routing & territory management. See https://www.opendoorlogistics.com/data/ (accessed on 18 December 2020).

29. Prener C, Revord C, Fox B. 2020 *Areal: Areal Weighted Interpolation*. See https://CRAN.R-project.org/package=areal.

30. R Core Team. 2017 *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

31. Lacasa L, Challen R, Brooks-Pollock E, Danon L. 2020 A flexible load sharing system optimising ICU demand in the context of COVID-19 pandemic.

32. Gregory S. 2010 Finding overlapping communities in networks by label propagation. *New J. Phys.* **12**, 103018. (doi:10.1088/1367-2630/12/10/103018)

33. Sun H-L, Huang J-B, Tian Y-Q, Song Q-B, Liu H-L. 2015 Detecting overlapping communities in networks via dominant label propagation. *Chinese Phys. B* **24**, 018703. (doi:10.1088/1674-1056/24/1/018703)

Supplementary material - Estimating surge hospital capacity in Britain during the COVID-19 pandemic

Identifying a set of capacity data for the NHS proved complex. After several attempts to integrate data from various sources, we ultimately performed a manual curation of the sources listed below, with gaps or inconsistencies filled in by consultation with the relevant hospital's website. The resulting list is a snapshot in time of capacity and not representative of up to date practice. During the source of the COVID-19 pandemic a small number of NHS trusts merged which had to be adjusted for. There are also significant limitations due to the different ways the devolved administrations of the UK (England, Wales, Scotland and Northern Ireland) reported situation report of bed capacity durign the pandemic, which meant only England and Wales hospitals has assessments of surge capacity, and we had no reliable information about Northern Ireland at all, and hence it was excluded. This does not significantly alter our conclusions here about the nature of the algortihm, but should be bourne in mind, if the data set is to be used for other purposes.

NHS and Trust GIS locations (England):

- https://www.nhs.uk/about-us/nhs-website-datasets/
- Lists of independent and NHS hospitals and trusts with location data
- public

NHS Trusts (England)

- https://www.nhs.uk/ServiceDirectories/Pages/NHSTrustListing.aspx
- Lists of NHS trusts and locations (as postcode) with information about services offered and hospital sites
- public

Beds open - NHS England:

- https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-overnight/
- https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-day-only/
- Information at an NHS trusts level on hospital beds and icu beds available
- public

Critical care capacity in England (pre-pandemic):

- https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/critical-care-bed-capacity-and-urgent-operations-cancelled-2019-20-data/
- Prepandemic NHS trust bed and ICU capacity
- public

Wales:

Average daily beds by site:

- https://statswales.gov.wales/v/Hg4K
- Prepandemic ICU and general bed availability
- public

Scotland:

Annual trends in available beds:

- https://www.isdscotland.org/Health-Topics/Hospital-Care/Publications/data-tables2017.asp?id=2494#2494
- Prepandemic Hospital and ICU bed capacity
- public

Sitrep (Situation reports) data:

England:

- filename: Covid sitrep report incl CIC 20200408 FINAL.xlsx
- Acute and ICU beds available in England at site level
- ICU (SIT032) and HDU (SIT033) beds available - many data quality issues and missing trusts
- restricted

Wales:

- filename: NHSWalesCovid19Sitrep-20200408.csv
- Acute and ICU beds available in Wales
- restricted

N.B. No sitrep data for Scotland or for Northern Ireland