

© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2009.01069.x
METHODS ARTICLE

Methods Articles

Identification of Hospital Catchment Areas Using Clustering: An Example from the NHS

Stuart John Gilmour

Objective. To develop a method of hospital market area identification using multivariate data, and compare it with existing standard methods.

Data Sources. Hospital Episode Statistics, a secondary dataset of admissions data from all hospitals in England, between April 2005 and March 2006.

Study Design. Seven criteria for catchment area definition were proposed. *K*-means clustering was used on several variables describing the relationship between hospitals and local authority districts (LADs) to enable the placement of every LAD into or out of the catchment area for every hospital. Principal component analysis confirmed the statistical robustness of the method, and the method was compared against existing methods using the seven criteria.

Principal Findings. Existing methods for identifying catchment areas do not capture desirable properties of a hospital market area. Catchment areas identified using *K*-means clustering are superior to those identified using existing Marginal methods against these criteria and are also statistically robust.

Conclusions. *K*-means clustering uses multivariate data on the relationship between hospitals and geographical units to define catchment areas that are both statistically robust and more informative than those obtained from existing methods.

Key Words. Hospital catchment area, clustering, principal component analysis, power law, Marginal method

The identification of hospital catchment areas is an important challenge in health services research (Garnick et al. 1987). Hospitals receive patients from geographical units, some of which represent a large proportion of the hospital's total activity, and some of which are physically far removed from the hospital. Classifying these geographical units into or out of a catchment or market area is necessary to define the major contributors to a hospital's business (Folland 1983), properties of a hospital's activity (Ashton and Press 1997; Bissegger 2006), and a population baseline from which to calculate expected

burden of disease and associated admission rates (Glynn et al. 1993; Wennberg et al. 2004). Questions about the relative rates of admission to hospital, differences in mortality rates between hospitals, and changes in hospital usage over time cannot be answered without knowledge of the demography of the areas a hospital is expected to serve. In the United States, catchment areas have also been used in antitrust legislation (Bissegger 2006), and their definition can have a significant impact on hospital business.

The British National Health Service (NHS) is a comprehensive health service available to all on the basis of clinical need rather than the ability to pay (Department of Health 2009). The NHS uses a “gatekeeper” model of primary care in which general practitioners (GPs) provide services to geographically local and refer these patients primarily to geographically local hospitals. Recently, enhanced patient choice and the introduction of competition have opened hospitals to geographically dispersed patients. Providers of secondary care services (“providers”) are being split from their fund-holders (“commissioners”), and both providers and commissioners need to assess aspects of their services that depend on definition of catchment areas. External monitoring agencies such as Monitor and the Competition and Cooperation Panel need clear definitions of market areas to assess whether competition is serving local populations well. Private providers of information services have also begun to explore the definition of hospital catchments (Ernst and Young 2009).

It is in this context that this article compares common methods used to define catchment areas outside of the United Kingdom with improved statistically defensible method based on *K*-means clustering, which is applicable in a wide variety of geographical and market contexts.

METHODS USED TO CALCULATE CATCHMENT AREAS

Early methods for the calculation of catchment areas used geographical boundaries, typically a circle of a specified radius around the physical location of the hospital (Phibbs and Robinson 1993). These rules have been superseded by patient flow methods (Baker 2001) that assign geographical units to a hospital catchment area if the proportion of the hospital’s total activity from that

Address correspondence to Stuart John Gilmour, B.Sc. (Hons.), M.Sc. (Statistics), M.P.H., Research Fellow, The King’s Fund, 11-13 Cavendish Square, London W1G 0AN, United Kingdom, e-mail: s.gilmour@kingsfund.org.uk

geographical unit is above some threshold value or *margin*. For example, geographical units providing > 0.5 percent of total hospital admissions might be assigned to the hospital catchment area. Alternative marginal rules analyze the proportion of a geographical unit's admissions that are to a given hospital (Basu 1994). Various embellishments on these *marginal rules* have been proposed, primarily with the purpose of eliminating units that contribute unusually low numbers to hospital activity while ensuring a minimum (or maximum) proportion of activity is represented by the catchment area. These rules have been shown to lead to very similar calculations of market concentration indices for a given marginal rule across a wide range of threshold values (Zwanziger, Melnick, and Mann 1990), but limited attention has been paid to the identification of market areas for defining burden of disease or investigating the statistical validity of the methods.

Many problems with marginal rules arise from the arbitrary nature of the margin and the limited amount of information used to define the rule. Values of the margin should depend on the scale of the hospital's activity, the number of geographical units served, or their size. However, even were the margin to be set for a class of hospitals of similar size and activity, its threshold value remains arbitrary. Existing marginal rules do not specify how the margin should vary between geographical units of differing detail—for example, different sizes of small area—or even whether such variation is necessary. The marginal rules also do not use the full range of information available about the relationship between hospitals and the areas they serve, such as patient flow *and* geographical information. This can lead to perverse situations where, for example, small and low-populated regions located very close to large teaching hospitals are not considered to be part of those hospitals' catchment areas.

In this paper, we propose that a method for calculating hospital catchments should do the following (Adams and Wright 1991):

1. Capture a significant portion of the hospital's activity.
2. Exclude areas whose contribution to hospital activity probably represents random variation.
3. Reflect geographical influences on hospital activity, such as transport links and physical barriers to access.
4. Reflect cultural and historical influences on hospital activity, such as consultant-to-consultant relationships.
5. Differ by specialty or type of attendance.
6. Increase with hospital size.
7. Change seasonally for hospitals in holiday or seasonal working areas.

The particular weighting given to the seven criteria is likely to differ according to the purpose of the catchment area definition and the system within which it operates, so in this paper we consider these criteria as they apply to the NHS. Given that referral patterns in the NHS were initially geographically localized and developed through relations between consultants, GPs, and local authorities, any method for calculating catchment areas must reflect points 3 and 4, though it may not use information about them directly, and it is unlikely that a catchment area defined on a single variable will represent these historical, cultural, and geographical phenomena well. In order to reflect criteria 3 and 4, geographical *and* patient flow information should be used, as should the information contained in the relationship between these variables. Criteria 5–7 require a flexible, data-driven method for assessing the relevance of the geographical units of interest. In different circumstances it may also be necessary to weight variables differently, another consideration better achieved with a data-driven method rather than a pre-conceived model. Criteria 1 and 2 will be in tension in any catchment area definition, and the use of well-founded data-driven statistical methods is most likely to resolve this tension successfully.

This paper describes a method for identifying catchment areas using *K*-means clustering in a multivariate dataset, bolstered statistically through an equivalent application of principal components analysis (PCA). This method reflects theoretical expectations about the desirable properties of a market area, varies automatically and flexibly with the size and distribution of a hospital's activity, uses a diversity of information about the relationship between hospitals and the geographical units they serve, incorporates geographical information, includes measures of uncertainty, provides a measure of capture of a geographical unit into a catchment area, and is statistically justifiable. It produces intuitively defensible catchment areas, provides a measure of catchment area inclusion that is amenable to linear regression for the study of changes in catchment inclusion over time or by properties of geographical units, and enables measures of uncertainty to be calculated.

The method is tested on data for all hospitals in the NHS and shown to produce market areas that match the proposed criteria. An example is presented for a specific hospital from the NHS. Little effort has been made to identify catchment areas for hospitals in the NHS, even though small area statistics are available and geographically definable catchment areas were a historical reality. It is hoped that this work will provide both an improvement in catchment area identification methods and catchment areas for every

hospital in the NHS. Datasets containing catchment area information can be obtained from the author.

METHOD

Hospitals in England are administered by Trusts, which usually manage several sites. Because in some instances the data on individual sites are not clearly recorded, this analysis was conducted at the level of the Trust. For every Trust a dataset of three variables was constructed, with one observation for each local authority district (LAD) that provided at least one patient admission to the hospital. The three variables are referred to as follows:

- Proportion of hospital admissions: the percentage of a hospital's admissions that came from each LAD.
- Proportion of LAD market: the percentage of an LAD's total admissions to all hospitals that occurred at a given hospital.
- Distance: the distance between a hospital and an LAD.

These are the three variables historically used for the construction of hospital market areas, and their combination in a multivariate model enables us to explore the interaction between these key measures of the relationship between a hospital and its tributary geographical units.

These data were extracted from the hospital episode statistics (HES) dataset for every Trust in England between April 2005 and March 2006. Attendances were calculated for all services (elective and nonelective) for every Trust, although more detailed analysis of some specialties was also conducted and is available from the author on request. Data on all services were used and presented here for testing because, as total numbers of attendances increase, comparisons between methods become more accurate. Unstandardized population data were used so as to perform the simplest possible comparison between the proposed and existing methods. However, nothing precludes the use of standardized data, and the method has been developed to be applicable to demographic subgroups without the need to change any threshold values in the underlying model.

Physical coordinates for every LAD and Trust were obtained from the Office of National Statistics and used to calculate an estimate of the distance between the Trust and the LAD. All variables were variance stabilized by taking natural logarithms, and centered and standardized so that they lay on the same scale. Variance stabilization in this data is essential, because a small

number of extreme values will otherwise carry excessive weight, and is routine in multivariate analysis (Johnson and Wichern 2002). For all three variables, the unstabilized data were highly skewed. The population distribution of LADs is largely confined to a small range of values, and the larger LADs tend to be urban, though there is no systematic pattern to the distribution of LAD populations. Admissions from individual LADs can be shown to have a strong linear relationship with LAD population (data available on request).

The proportion of admissions and the proportion of LAD market were calculated separately for each Trust, giving a separate dataset of three variables for each Trust. Separately for each Trust, the transformed data were divided into two clusters using *K*-means clustering, a well-established method for dividing a multivariate dataset into clusters based on the distance between observations, as follows (MacQueen 1967):

1. Two cluster centers are chosen arbitrarily.
2. Each observation is assigned into the cluster whose center it lies closest to.
3. The center of the cluster formed by this assignment is recalculated.
4. The process is repeated until the cluster assignments cease to change.

In *K*-means clustering, “distance” refers to the distance between observations in a Euclidean space with dimensions given by the variables of interest in the dataset—one of which, in this application, is a variable measuring distance from LAD to Trust.

Division of a multivariate dataset into two clusters is mathematically equivalent to dividing it into two groups according to the sign of the first principal component (PC1) (Ding and He 2004) from PCA. PCA is a standard method of data reduction for multivariate data (Johnson and Wichern 2002), with several desirable properties:

- Principal components are calculated from the covariance matrix of the data, and they use all available information about the relationship between the variables in the dataset.
- PC1 explains the largest amount of variance in the data.
- Principal components are closer to normally distributed than the original data.

These properties of PCA motivate its use as an adjunct to *K*-means clustering in statistical tests of the parameters of the catchment area. Because a 2-means clustering is mathematically equivalent to PCA, and the proportion

of variance explained by PC1 can be calculated, PCA enables explicit calculation of the proportion of variance captured by the catchment area clustering. The equivalence of PC1 and 2-means clustering also enables a measure of market penetration for the 2-means clustering to be calculated from PCA. This is possible because PC1 has a value for every LAD in every catchment area, with the value of 0 as the cut-off for inclusion in the catchment area. Because the LADs, which are marginal for inclusion into the cluster will tend to have a value of PC1 that lies closer to zero, judgments can be made about the extent to which a Trust has “captured” a market area. Data on “capture” of an area are not presented in this paper, which focuses on testing the comparability of the PC1/*K*-means equivalence.

The *K*-means method was compared with a straight marginal rule with 0.5 percent threshold. Under this rule, any LAD that represents more than 0.5 percent of Trust admissions is assigned into the Trust’s catchment area, and all other LADs are excluded. This rule has been recommended in other catchment area studies (Zwanziger, Melnick, and Mann 1990).

Comparison of K-Means and Marginal Methods

The *K*-means and Straight Marginal methods were compared on three different measures related to criterion 6, viz. that there should, in general, be some relationship between Trust size and the size of the catchment area. For both methods, the following statistics were calculated:

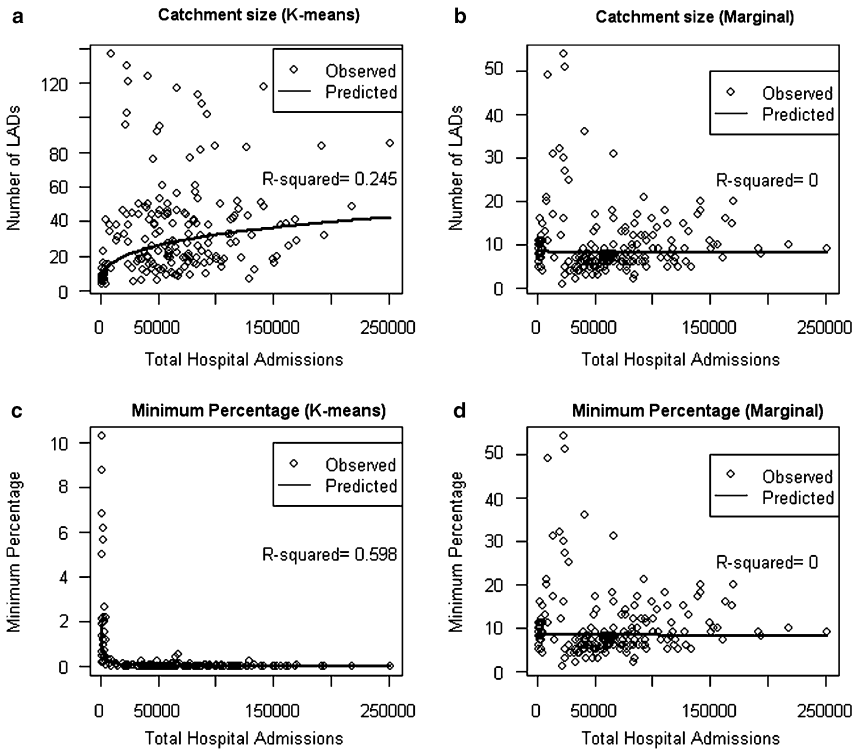
- Size: the number of LADs included in the catchment.
- The smallest proportion of Trust business captured by an LAD included in the catchment area.
- The distance to the furthest LAD included in the catchment area.

To test assumption 6 rigorously, power law models were fit to these measures as a function of total Trust activity. The fitted lines from these models and the total variance in the log-log data explained by these models are plotted for the first two indicators (Figure 1).

Both methods were compared on the basis of criterion 1, viz. that they captured significant portions of the Trust’s overall admissions, by graphical comparison of the percentage of all admissions to each Trust which came from the LADs included in the catchment area (Figure 2).

The two methods were compared directly, by plotting the percentage of LADs assigned into the catchment by the *K*-means method, which were also assigned into the catchment by the Marginal method (Figure 3).

Figure 1: Catchment Area Size and Minimum Percentage for Inclusion in Catchment Area by Total Hospital Admissions



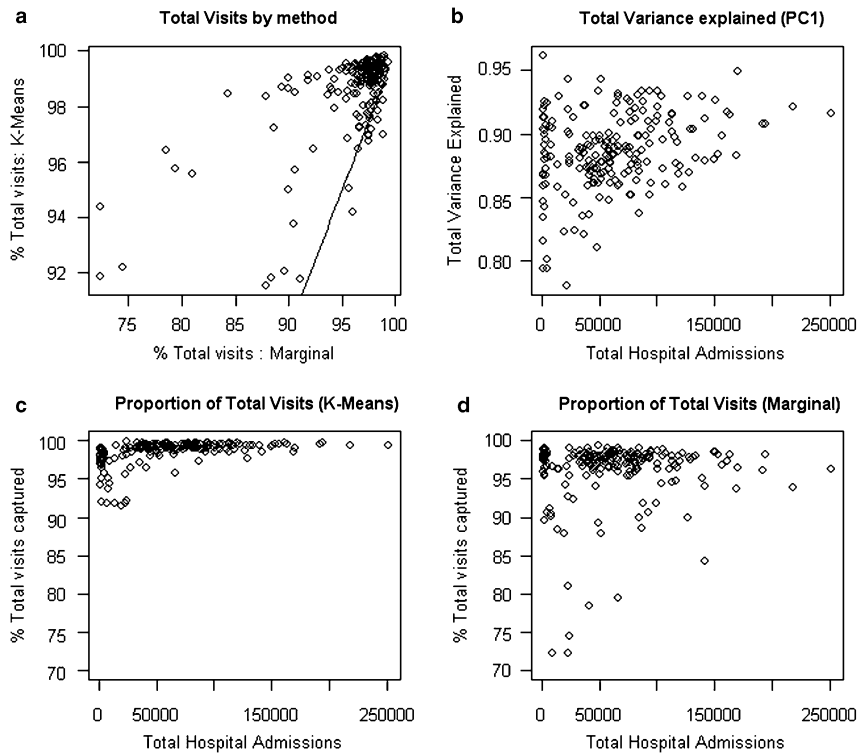
Note. Plots of number of LADs included in catchment area (a and b) and minimum threshold percentage for inclusion in the catchment area (c and d) for the *K*-means and the Marginal methods. Fitted lines from the power law (in the log-log transformed data) are plotted on every graph, and the R^2 value for each model is shown on the chart.

Testing Equivalence of K-Means and PCA Clusters

The equivalence of *K*-means clustering and PC1 clustering is only guaranteed in the population case, and in practice there is some sampling error. To test the level of agreement, two additional statistics were calculated based on the PCA approximation:

- Proportion of total variance explained by the PC1, which shows approximately how much of the total variance in the three variables is explained by the PCA clustering, and

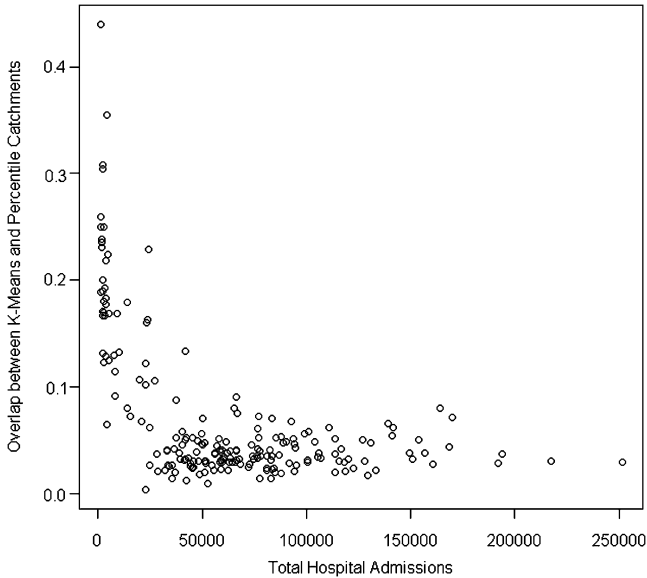
Figure 2: Proportion of Total Visits and Variance Included in the Catchment Area by Model Type



Note. Tests of relative proportions of all hospital admissions or variance explained by the two methods. (a) plots the proportion of the total hospital admissions that comes from the catchment area for the two methods. The line in this figure indicates the points, where the *K*-means and Marginal methods capture the same proportion of total hospital admissions. For all hospitals to the left of this line, the *K*-means method captures a higher proportion of admissions than the Marginal; for hospitals to the right of the line, the opposite applies. (b) shows the proportion of total variance that is explained by PC1. (c) and (d) plot the two axes of (a) separately, against total hospital admissions, with *y*-axes on the same scale. Neither plot is well explained by a power law.

- Percentage of LADs assigned into the catchment by both *K*-means and PC1, which indicates the extent to which the information gained from PC1 can be used to describe the *K*-means clustering.

Figure 3: Proportion of LADs Assigned into Market Areas by *K*-Means and Marginal Methods



Finally, the cluster assignments for University College Hospital London (UCLH) are plotted for both *K*-means clustering and a straight marginal rule of 0.5 percent as an example.

RESULTS

Cluster assignments were produced for 201 Trusts. These Trusts serviced between 17 and 340 LADs, and an average of 193. The comparison of *K*-means and Marginal methods is tested in two sections: first, the relationship between catchment area properties and Trust admissions is tested with respect to assumption 6; secondly, the agreement between *K*-means and Marginal methods is tested. The equivalence of *K*-means and PCA catchment area calculations is then presented. Finally, the example of UCLH is described.

Throughout the results, the term *Marginal method* should be taken to mean an application of the straight Marginal method with threshold of 0.5 percent.

Catchment Area Properties and Trust Admissions

The number of LADs included in Trust catchment area is plotted against total Trust admissions in Figures 1a and b. A power law model fits the *K*-means catchment area, with the number of areas included in the catchment area increasing as approximately the cube root of Trust admissions. This power law explains 25 percent of the variance, compared with 0 percent for the Marginal method. This supports the assumption that the catchment area of a Trust should increase with size; but Figure 1 does appear to identify a group of outlier Trusts that follow a different relationship. This group of Trusts included notable specialty providers such as Moorfields Eye Hospital and the Royal National Orthopaedic trust, suggesting that this group of outliers represents Trusts that provide specialty services across a very wide area. These results give further support for the validity of this method in linking catchment areas to properties of Trusts.

The minimum percentage of Trust admissions captured by catchment area LADs is plotted against total Trust admissions in Figures 1c and d. This variable shows the percentage of Trust admissions captured by the LAD in the catchment that provided the smallest proportion of Trust admissions. The *K*-means method follows a clear power law, explaining 59 percent of total variance. The minimum percentage of Trust admissions required for an LAD to be included in the *K*-means catchment area declines as the inverse of Trust admissions to the power of approximately 1.1. By contrast, the minimum percentage for inclusion in the Marginal method is not explained by any power law, reflecting the threshold nature of the marginal rule used. This is strong evidence that the threshold for inclusion of an area needs to be sensitive to the distribution of the data being used, and it cannot be expected to be fixed across different Trusts of varying size and characteristics.

Figure 2a plots the total percentage of Trust admissions included in the catchment area for the *K*-means method, plotted against the same statistic for the Marginal method (Figure 2a), and Figures 2c and d show this percentage as a function of total Trust admissions for the *K*-means and Marginal methods, respectively. Figure 2b shows the proportion of total variance in Trust admissions contributed by each LAD that is explained by the PC1 market area assignment.

It is clear from Figure 2 that a higher proportion of all visits is included in the catchment area by *K*-means than by the Marginal method, as illustrated by the red line in Figure 2a; only points lying below the red line capture the situation, where the Marginal method includes more visits. Note that sug-

gested embellishments on the marginal rule, such as the maximum marginal rule, will not change the structure of the identified catchment area. In general, maximum marginal rules aim to capture a minimum percentage of total Trust admissions of 70–80 percent, which threshold is exceeded by the vast majority of Trusts under the inclusion threshold (0.5 percent) used in this paper. Figure 2 shows clearly that the *K*-means method both captures a large proportion of the total Trust admissions in the catchment area, and a large proportion of the variance of the LAD data from which the catchment area is originally composed.

The furthest distance to an LAD that was included in the catchment area was also analyzed, though the plots are not shown here. A power law relationship exists between the maximum distance to an included LAD and total admissions to the Trust, explaining 20 percent of the variance in the log of distance for the *K*-means method, but none of the variance for the Marginal method. Maximum distance increases approximately as the fifth root of Trust admissions.

Agreement of K-Means and Marginal Methods

Figure 3 shows the proportion of LADs assigned to the catchment area by the *K*-means and the Marginal methods, plotted as a function of total Trust admissions. This plot reflects the findings of Figures 1–3, that catchment size does not depend on Trust size for the Marginal method. For larger Trusts the *K*-means method expands the catchment size, but it remains the same or even reduces under the Marginal method, leading to a smaller and smaller proportion of overlap. In all cases, the Marginal method catchment area lies almost entirely within the *K*-means catchment.

The *K*-means method includes more areas in the catchment area, as can be seen in Figure 4, and this explains the higher proportion of total visits displayed in Figure 2. The extra areas included in the catchment by *K*-means do not just represent a larger catchment area; they provide additional shape and structure that may be consistent with the real underlying factors driving use of hospital services.

Confirmation of Equivalence of PCA and K-Means Clustering

Finally, we confirm the overlap between catchment areas identified by the *K*-means methods and catchment areas derived from PC1. Because statistical uncertainty means that *K*-means and PCA definitions of catchment areas do not have to overlap, the proportion of agreement between the two methods

Figure 4: University College London Hospital, *K*-Means and Marginal Methods



was calculated as the proportion of all LADs in which catchment area assignment agreed, so every LAD that was assigned into *or* out of the catchment area by both methods was considered to represent equivalence in catchment area assignment. Agreement between *K*-means and PCA methods for defining catchment areas is good (average 82 percent, range 62–100 percent), and independent of the number of LADs from which the Trust draws patients, although there is some evidence that variance is larger for Trusts receiving patients from larger numbers of LADs. This confirms that PCA provides a good approximation to the *K*-means clustered catchment area and that the values of PC1 can be expected, in general, to reflect the underlying *K*-means clustering. For example, values of PC1 close to 0 may represent higher uncertainty about the correct assignment of an LAD into or out of a catchment area, while large absolute values of PC1 represent a clear assignment. This may be useful in plotting changes in market penetration over time. Confidence intervals for any value of PC1 for a given LAD, or for cluster centers, can also be calculated.

Practical Example: UCLH

The catchment area for UCLH is presented in Figure 4. The catchment from the *K*-means method is shown in the top figure and the Marginal method is on the bottom. In both charts, the blue-colored regions lie outside the catchment, while the purple-colored regions lie inside the catchment.

The *K*-means method includes a much larger number of LADs and describes a shape that represents transport routes to central London. The shape is also not symmetric about the hospital location, while that for the Marginal method is. The Marginal method appears to estimate only the core part of the UCLH catchment area, while the *K*-means method constructs a shaped catchment area representing greater London and some regions on convenient transport routes. Catchment areas for specialties, or divided up by elective/nonelective services, appear different to those shown here. A full dataset of catchment areas for every hospital for selected specialties or for all services is available from the author.

DISCUSSION

Existing marginal rules for establishing hospital catchment areas have several drawbacks, caused by the arbitrary nature of the margin and the use only of

univariate data. This paper presents a multivariate statistical approach with several advantages:

1. *Data driven*: geographical units are included in the catchment area empirically from the data.
2. *Multivariate*: Multiple variables are incorporated into the catchment area assignment.
3. *Statistically robust*: the method enables calculation of confidence intervals, the proportion of variance explained by the catchment area assignment, and the degree of market penetration of any area by any hospital.
4. *Geographic and patient-flow driven*: the method combines two historical methods for identifying catchment areas.
5. *Extensible*: the method can be extended to include new variables, including variables that are not amenable to marginal rules (such as measures of market concentration).
6. *Flexible*: the method can be extended to GP surgeries, census output areas, health administrative areas such as primary care trusts (PCTs), or small areas in other geographical or national settings, without any fundamental change in the method, and without any risk of misspecification of a marginal cut-off.
7. *Customizable*: the model can be applied to demographic subgroups and enables hospitals to identify subpopulations in which they perform poorly and possibly even to model demographic change, without assumptions about starting parameters such as the threshold for inclusion of an area on a variable of interest.

The statistical robustness of catchment areas constructed using K -means clustering can be investigated rigorously through the use of PC1, and confidence regions placed around their cluster centers. Values lying in the confidence intervals of both cluster centers (where the confidence intervals overlap each other but not the cluster centers) can be defined as uncertain in assignment. This also allows quantification of uncertainty in catchment definitions (if the confidence region around one cluster center overlaps the other), which may be possible if the small areas or specialties used for analysis provide very small numbers. Current catchment area identification methods do not have any associated tests of statistical robustness or measures of their uncertainty. All catchment area methods are *unsupervised learning* methods, which are generally accepted as being difficult to test statistically (Hastie, Tibshirani,

and Friedman 2003), but the method proposed here uses two classical methods of unsupervised learning whose statistical validity is well accepted.

The profile of an example London Trust shows a very different and much more interesting structure under the K -means method than the Marginal method. It is a simple task to define the population serviced by this hospital, its age and sex structure, and the expected prevalence of disease, enabling estimation of the expected level of health service needs for comparison with observed demand or activity statistics. It is also possible to calculate catchment areas that vary seasonally, or by specialty, subpopulation, or service type, and to use the observed values of the PC1 to forecast changes in catchment areas over time. The method used here is also suitable for analyzing subgroups of LADs or Trusts, and relationships between covariates describing LADs and their inclusion or exclusion from the catchment areas. For space reasons such analyses have not been presented here.

In the context of the NHS, the introduction of competition into hospital services makes the identification of catchment areas essential for providers, commissioners, regulatory bodies, and corporate providers of information services. The method presented here is sufficiently flexible and versatile to be applied to any health care market, however, so although it has been tested using NHS data it should be equally applicable in other health economies, all of which depend on catchment area definitions for service planning and analysis.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: I would like to acknowledge the assistance of Dr. Francesca Frosini, Senior Analyst, the King's Fund, for helpful suggestions regarding the content of the paper.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Adams, E. K., and G. E. Wright. 1991. "Hospital Choice of Medicare Beneficiaries in a Rural Market: Why Not the Closest?" *Journal of Rural Health* 7 (1): 134–52.
- Ashton, T., and D. Press. 1997. "Market Concentration in Secondary Health Services under a Purchaser-Provider Split: The New Zealand Experience." *Health Economics* 6: 43–56.

- Baker, L. C. 2001. "Measuring Competition in Health Care Markets." *Health Services Research* 36 (1): 223–52.
- Basu, J. 1994. "An Analysis of Market Shares of Maryland Hospitals in Their Service Areas." *Journal of Health & Social Policy* 6 (1): 71–85.
- Bissegger, M. R. 2006. "The Evanston Initial Decision: Is There a Future for Patient Flow Analysis?" *Journal of Health Law* 39 (1): 143–59.
- Department of Health 2009. *The NHS Constitution: The NHS Belongs to Us All*. London: Department of Health.
- Ding, C., and X. He. 2004. "K-Means Clustering via Principal Component Analysis." In Carla E. Brody (Ed.), *Proceedings of the International Conference on Machine Learning (ICML 2004)*, pp. 225–32. Banff, Canada: ICML.
- Ernst & Young 2009. *Understanding Health Care Markets: A PCT Guide to Market Analysis and Market Management*. London: Ernst & Young.
- Folland, S. T. 1983. "Predicting Hospital Market Shares." *Inquiry* 20: 34–44.
- Garnick, D. W., H. S. Luft, J. C. Robinson, and J. Tetreault. 1987. "Appropriate Measures of Hospital Market Areas." *Health Services Research* 22 (1): 69–89.
- Glynn, R. J., T. A. Stukel, S. M. Sharp, T. A. Bubolz, J. L. Freeman, and E. S. Fisher. 1993. "Estimating the Variance of Standardised Rates of Recurrent Events, with Application to Hospitalizations among the Elderly in New England." *American Journal of Epidemiology* 137 (7): 776–86.
- Hastie, T., R. Tibshirani, and J. Friedman. 2003. *The Elements of Statistical Learning*. 2d Edition. New York: Springer.
- Johnson, R. A., and D. W. Wichern. 2002. *Applied Multivariate Statistical Analysis*. 5th Edition. New Jersey: Pearson Education.
- MacQueen, J. B. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–97. Berkeley, CA: University of California Press.
- Phibbs, C. S., and J. C. Robinson. 1993. "A Variable-Radius Measure of Local Hospital Market Structure." *Health Services Research* 28 (3): 313–24.
- Wennberg, J. E., E. S. Fisher, T. A. Stukel, J. S. Skinner, S. M. Sharp, and K. K. Bronner. 2004. "Use of Hospitals, Physician Visits, and Hospice Care During Last Six Months of Life among Cohorts Loyal to Highly Respected Hospitals in the United States." *British Medical Journal* 328: 607–11.
- Zwanziger, J., G. A. Melnick, and J. M. Mann. 1990. "Measures of Hospital Market Structure: A Review of the Alternatives and a Proposed Approach." *Socio-Economic Planning Science* 24 (2): 81–95.