

The New Mexico Mammography Project: Using GIS to Determine Geographic Variation in Mammography Utilization

Andrew M Amir-Fazli,* Patricia M Stauber, Meg Adams-Cameron, Charles R Key
New Mexico Tumor Registry, Cancer Research and Treatment Center, University of New Mexico,
Albuquerque, NM

Abstract

The New Mexico Mammography Project (NMMP) is establishing a population-based mammography registry for the state of New Mexico. One of the aims of this project is to examine the effectiveness of mammography screening in a community setting. In order for mammography screening to achieve the degree of reduction in mortality from breast cancer that has been demonstrated in randomized control trials, women at risk must undergo screening on a regular schedule. According to a previous analysis of over 135,000 women's records in the NMMP database, more than 60% of all women had at least one mammogram between 1992 and 1995, but only 22% had three exams in those four years. There are many reasons for lack of compliance. One that few studies have examined is the effect of travel distance to a mammography facility. Information available in the NMMP database enabled us to determine, for each examination, where the woman being examined lived and where the examining facility was. Over 80% of recent addresses were geocoded to the census tract level and 99% to the zip code level. Geographic information system (GIS) technology, specifically ArcView 3 software with the Network Analyst extension, was used to calculate the distances women drove to mammography facilities. Mammography screening rates for various areas of the state were then calculated and analyzed (using US census data) for differences by driving distance versus age, rural or urban status, education, and household income. The results of this analysis are not yet available; this paper presents a discussion of the issues involved and the problems encountered in using GIS methodology with registry data.

Keywords: health care access, mammography, driving distance, geocoding

Introduction

The New Mexico Mammography Project (NMMP) is establishing a population-based mammography registry for the state of New Mexico (1). One of the aims of this project is to examine the effectiveness of mammography screening in a community setting. In order for mammography screening to achieve the degree of reduction in mortality from breast cancer that has been demonstrated in randomized control trials, women at risk must undergo screening on a regular schedule. According to a previous analysis of over 135,000 women's records in the NMMP database, more than 60% of all women

* Andrew M Amir-Fazli, New Mexico Tumor Registry, Cancer Research and Treatment Center, University of New Mexico, 2325 Camino de Salud NE, Albuquerque, NM 87131 USA; (p) 505-272-8575; (f) 505-272-8572; E-mail: aamirf@nmtr.unm.edu

had at least one mammogram between 1992 and 1995, but only 22% had three exams in those four years. There are many reasons for lack of compliance. One that few studies have examined is the effect of travel distance to a mammography facility. In 1997, the New Mexico Tumor Registry (NMTR) at the University of New Mexico Cancer Research and Treatment Center began a SEER (Surveillance, Epidemiology, and End Results) Special Study for the National Cancer Institute, "Geographic Variation in Breast Cancer Treatment and Mammography Use." An objective of this study was to use geographic information system (GIS) technology to determine the distances from patients' residences to treatment or diagnostic facilities, and analyze how these distances affect treatment choice or service utilization.

Overview

To compute mammography utilization rates, NMTR selected records from the NMMP database for all women age 40 and older who had a mammogram in 1994 or 1995 at any of the screening facilities in the five-county area surrounding Albuquerque, the largest city in New Mexico. At the time of the study, this was the only area in New Mexico for which data on all mammograms (more than 95%) had been collected. Mammograms were grouped into series called mammographic events; an initial mammogram and immediate follow-up examinations (all examinations within 90 days) were considered one mammographic event. Information in the NMMP database was used to determine where the women lived at the time of their examinations and locate the mammography facilities where the examinations were performed. For most areas, over 80% of recent addresses were geocoded to the street address (census tract) level and 99% to the zip code level. Age-adjusted rates of mammography utilization (the number of mammograms per 100 women per year) were calculated. The population numbers for the denominators were taken from the 1990 US Census Bureau zip code files.

An initial statewide study, "Geographic Variation in Breast Cancer Treatment," used a smaller number of cases and facilities from the years 1994 and 1995 to develop and test GIS methods. GIS technology was used to calculate the distances women drove to radiation treatment centers and mammography facilities. Once methods were established, they were applied to the larger mammography database so that mammography screening rates could be calculated for various areas of the state. US census data were used to analyze the rates in these areas for differences by driving distance versus age, rural or urban status, education, and household income. The hypotheses for the study were that rates are lower in areas with the greatest distance to travel to receive a mammogram, rates are lower in rural areas than in urban areas, and rates are lower in areas with lower educational levels or lower household income. As of this writing, the final analysis phase of the study is not complete; this paper presents results of and problems arising from using GIS.

Using GIS

Methods

In order to allow analysis of geographic variation in the NMMP data collected, it was necessary to develop GIS capability at NMTR. The GIS was created specifically for two

main operations, geocoding and routing. Geocoding is the process of assigning an absolute location (in this case, latitude and longitude coordinates) to a geographic feature referenced by a relative location (such as a street address or zip code). Once facility and patient locations were determined by geocoding, a Euclidean (or straight line) distance could be calculated between patients and facilities and a measure of geographic variation, such as a distance to the closest facility, could be determined. A GIS, however, is capable of more sophisticated routing analysis, determining a shortest distance along available road networks. Because it is not possible in most cases to travel in a straight line in New Mexico, a GIS network analysis was used to compute driving distance (and to consider driving time) as a more realistic measure of geographic access.

The GIS installed at NMTR for this study was ArcView 3.0a (ESRI, Redlands, CA), with the Network Analyst extension. This software was available at low cost through the University of New Mexico site license and included the ESRI StreetMap product as a data resource. The system was installed on a generally available personal computer, a PC running Microsoft Windows 95. The PC was equipped with a Pentium Pro 233 MHz processor with 64 Mb RAM, an 8 Gb hard drive, and a 21-inch display. This configuration was adequate to the task, although some operations took several hours.

Geocoding

Geocoding involved many steps, including obtaining addresses of facilities and patients, standardizing addresses, obtaining a street reference file, configuring and running the geocoding process, and mapping the resulting locations. Where a patient's street address could not be successfully geocoded, the patient's zip code was used to determine a location of residence. To use zip codes required an additional step, assigning a point location for each zip code area. Geocoding by zip code allowed all patients to be assigned a location.

A file of 12 functioning radiation treatment facilities and their street addresses was created from NMTR's facility records and from consultation with radiation treatment personnel. All radiation treatment facilities were successfully geocoded to a street address. Addresses of mammography screening facilities were obtained from the Food and Drug Administration (FDA) Web site listing facilities certified by the Mammography Quality Standards Act. To ensure completeness, the FDA list was compared with the facility list from which patient records were obtained. A missing facility address for the Veterans' Administration Hospital/Kirtland Air Force Base was then added, making a total of 57 geocodable facilities. After mapping these locations, creating a five-county service area, and eliminating facilities located outside the service area, the number of facilities in the study was reduced to 21.

The addresses of radiation treatment cases were obtained from NMTR files, in which an address at time of diagnosis is routinely recorded in the course of entering cancer data. The addresses of mammography cases were obtained via records collected by the NMMP from screening facilities. In these mammography records, a site and subsite code were intended to identify the facility where the screening was performed. In the 1994–1995 patient records, the subsite code is currently missing for almost all cases (although it is believed this information can still be obtained). Without a subsite code, it cannot be known whether a patient record shows that the screening was performed in an outlying clinic and later read at a main facility, or whether the record shows that the screening was actually performed at a main facility. A distribution of the geocoded

1994–1995 mammography patient addresses shows that at least 16% of patient residences were outside the five-county service area defined for our mammography facilities. These records were excluded because it has yet to be determined whether these patients actually traveled great distances—away from closer mammography facilities—to use the facilities in the center of the state.

The geocoding component of a GIS is most successful when street addresses follow a standardized format. NMTR uses address-editing routines that conform to ArcView's expectations. A separate step was not required to reformat addresses recorded by NMTR. Records from mammography screening facilities varied in format quality. In the future, additional address-formatting software could be used to process these records. This may improve the geocoding match rates.

In order to geocode and perform routing analysis, a reference dataset of street locations and address ranges is required. This dataset is commonly called a street network file. The street network file used for this study was ESRI's StreetMap product, a dataset based on Geographic Data Technology's Dynamap/1000 dataset (GDT, Lebanon, NH), which in turn was constructed from the original US Census Bureau TIGER/Line files with enhancements and corrections. StreetMap was used to geocode facility and patient addresses directly. To be used as a street network file, portions of StreetMap were converted to the "shapefile" format used by ArcView. The reduced size of these StreetMap "regions" limited the area to be searched during network analysis and sped up processing on the PC. The large distances between radiation facilities made it possible to break up the state of New Mexico into separate StreetMap regions. These regions were overlaid on a map that contained zip code boundaries and major roads, and an analysis was performed on the zip codes that fell along the boundary of a region. These zip codes were then assigned to the region that had the nearest facility. For the five-county mammogram study, a region was constructed to cover the five counties and any outlying facilities that were near enough to the five-county boundary that they might be candidates for the nearest-facility calculation.

Late in the project time frame, TIGER/Line97 data became available for New Mexico. An evaluation was made to see if TIGER/Line97 would prove a better street network file resource than StreetMap. A software utility, TGR2SHP (GIS Tools, Knoxville, TN), was obtained and used to convert TIGER/Line97 files to a format compatible with the ArcView GIS software. The converted TIGER/Line97 files were then used to rerun both geocoding and street network conversion. A spot check of certain regions of the state indicated that the road classification problems (described below) in the StreetMap product were largely corrected. This indicated that the use of TIGER/Line97 could produce more realistic driving distances and could enable the use of driving time analysis. With properly classified roads, each type of road could be assigned an average speed limit and a time expended to traverse each road segment could be calculated and summed.

Geocoding with Tiger/Line97, however, produced significantly fewer matches and did not improve earlier methods. It appears that although TIGER/Line97 contains more recent information than the StreetMap product, which is based on earlier TIGER files, there are fewer total address ranges in TIGER/Line97's underlying database, resulting in a lower address match rate.

In New Mexico, many people, especially those living outside the main urban centers, use post office boxes, local road names, and rural route addresses. It was

recognized that not all patient records would have street addresses that would geocode. A secondary marker for location was needed, so all records had zip codes assigned. In order to geocode to a zip code, a reference dataset of zip codes was obtained. Two types of zip code datasets were originally available: a point coverage and a polygon coverage. These datasets were bundled along with StreetMap and were derived from information matching the census in the early 1990s. In the point coverage file were the zip code number, the associated postal name, a code indicating a type of zip code (whether an area or single site such as a PO box, office building, or entity such as a university), and an area in square miles. When the zip code points and polygons were overlaid on a map of New Mexico, the zip code points were found in most cases to be located at the geographic centers of each corresponding zip code polygon (area). A concern of the study was that certain zip codes in New Mexico cover very large geographical areas; for example, Roswell's zip area is over 5,000 square miles. In these cases, zip code would be a poor proxy for the location of a patient's actual residence.

When it was observed that many zip code points located at the centroid of a zip code area were positioned in roadless and uninhabited areas and far from the true population centers of the area, it was decided to use another zip code point file. The new file was obtained from the US Census Bureau LandView III system and contained zip code points established by the US Postal Service as of January 1, 1997. It was noted from a comparison of this file with the previous zip code file that the assignment of zip code numbers had not changed. The new zip code points represented, for the most part, actual post office locations (which as a general rule are close to the population centers of zip code areas). A plot of these new points overlaid on the older zip code points and New Mexico population centers showed a closer match between the new zip code points and New Mexico population centers. The new zip code file, however, assigned a number of less-used zip code points to the centroids of the counties. In the statewide radiation treatment study, any zip code point found in the patient data that was outside the mapped polygon boundary of the zip code was moved from the centroid of its county to the nearest major road segment within its proper zip code boundary. This was not done for the five-county screening study, although the newer zip code point file was used. A review of the five-county region map showed that very few of these county centroid points would have been moved any significant distance.

ArcView's geocoding function tags each address record processed as "matched" if ArcView can locate the address along a road segment in the street network file. All radiation treatment and mammography screening facilities were successfully matched to a street address, in some cases after the street address was corrected. Seventy-one percent (71%) of patient addresses for the radiation treatment study were geocoded to the street address. The original geocoding run had produced a match rate of only 66%. Unmatched records, however, were reprocessed interactively with the opportunity to make corrections and rerun the geocoding process. Corrections were made of obvious spelling errors, street numbers that extended existing ranges but did not cross zip code boundaries, non-standard address formats, and street suffixes that were different but still in the same zip code boundary. This type of "reject" processing was not done for the much larger file of mammography patient addresses.

Seventy-three percent (73%) of patient addresses for the mammography study were geocoded to the street address. Due to the number of missing address entries and the uneven quality of address information reported by mammography facilities, additional

address fields matched from the State of New Mexico Motor Vehicle Department (MVD) were appended to the patient records. Use of the MVD address fields consistently showed a higher geocoding match rate and was recommended over use of the original address fields.

Match rates for the mammography patient records were also increased by relaxing the matching sensitivity parameters of the ArcView geocoding process. This caused many "partial" matches to be assigned a map location and was similar in effect to manual corrections performed by interactive reprocessing of unmatched records. The lower-sensitivity settings accommodated errors such as spelling variations and address prefix or suffix differences (e.g., "Road" instead of "Drive"). Because zip code was used as a restriction, it was feasible to match additional records with the assurance that the resulting location would remain in the same zip code.

A concern of the study was the large discrepancy in geocoding rates between "urban" and "rural" counties. The two urbanized counties in the five-county study area had geocoding rates of 86% and 74%, while the three rural counties had rates of 56%, 49%, and 3%.

Whether or not geocoding to the street address was successful, all patient records were matched to mapped zip code points, so that these records could also be assigned driving distance values calculated from the zip code points.

Routing and Distance Calculation

Once facilities and patients and/or their associated zip code points were mapped, the shortest routes between facilities and patients were determined. For this analysis, Arcview's Network Analyst extension was used, specifically the FindClosestFac(ility) function. Because many possible routes can be mapped between patient and facility, the Network Analyst used a generally accepted heuristic algorithm to determine the shortest route. Driving distance was then calculated by summing the lengths of all road segments along the shortest path from the patient location to the facility location.

By default, ArcView reports the lengths of line segments in units of decimal degrees, which are not really units of linear measure. It was observed that summing these line lengths and using the UNITS.CONVERT function to change the total length into units of miles produced incorrect results. Therefore, a MILES field was added to any street network file used for reference and an Avenue (ArcView's programming language) script was modified to compute the proper length in miles for each road segment. Within this script, the function UNITS.CONVERTDECIMALDEGREES was called to correctly calculate a "great circle arc length" between the starting and ending point of each line segment. This is the proper calculation for line length on a latitude/longitude grid. This MILES field then was identified as the "cost" field to be used by Network Analyst in reporting the results of a shortest path analysis. The cost field is an additional value that can be summed and reported by Network Analyst.

An alternative considered was to calculate driving time. If the cost field could be recorded in terms of hours or minutes, then a driving time would be reported instead of a driving distance. This would have been useful, because the shortest distance computed is not always the shortest time. Many roads in New Mexico are barely passable at low speeds; while these roads may be direct paths, they are rarely driven, because a circuitous route via state or county roads is much faster. The Network Analyst function does not directly distinguish between fast and slow routes. When using an associated

cost field, however, the same effect can be achieved if there is a way of classifying roads in the street network file.

The line segment record in the data table associated with StreetMap retains the TIGER/Line census feature classification code, which serves as a road type attribute. The code scheme is as follows:

- A00 (found in the data but not a specified code)
- A11 through A18: primary roads and major highways
- A20 through A28: secondary roads and minor highways
- A30 through A38: connecting and county roads
- A40 through A48: city and neighborhood streets
- A50 through A73: service roads, 4WD trails, etc.

Computing different “costs” by road type would have allowed a driving time analysis. When the road network was mapped, however, it became evident that many thousands of unusable road segments were improperly classified as significant roads. A query and extract process selecting for the roads of types A11 through A38 was performed on the street network file created for each region. This process did filter out most non-drivable routes, and did leave intact the major routes.

Using the “filtered” street network file, a first pass was made to calculate paths from each zip code point in a region to the nearest facility. (Some zip code points were not located near enough to a line segment in the street network. These zip code points were moved small distances to the nearest line segment.) The resulting routes were checked individually by running Network Analyst in an interactive mode. This process created a driving distance value for each patient in each zip code area and allowed a visual check of the Network Analyst choice of paths. Several “broken” major routes were identified in this manner and repaired.

A limitation of Network Analyst as delivered is that it only computes one solution at a time from a user-selected menu option. This is impractical for studying thousands of cases. To remedy this, an Avenue script was written to “batch process” the cases in each region. This script identified, via a user input dialog, the street network, a file of “events” (in this case, either zip code points or street address points) and a key field identifier for each event, a file of facilities and associated key fields, and an output file in which to store results. The script then cycled through each event, performing the FindClosestFac function. At each cycle, an output record was stored indicating the event key, facility key, and minimum distance in cost units. John Fortney, Kathryn Rost, and James Warren of the University of Arkansas for Medical Sciences pioneered this approach and provided suggestions for our study (2).

The batch Network Analyst script was then used to process treatment patient records region by region. The results were appended to the treatment database for statistical analysis. Because both a zip code-derived distance and a street address-derived distance were stored with each patient record, a comparison of the two distances was made. A strong correlation ($R=.97132$) suggested that zip code distances might reasonably substitute for street address distances when processing the larger mammography records file. One hundred seven thousand eight hundred thirty-four (107,834) mammographic events were processed for the five-county study area for the two years studied. Zip code-derived distances were then also computed and assigned to each event record as a measure of quality control.

Further Study

This study will be expanded in future years, as more complete mammography records become available for more of New Mexico. It is believed that the effect of distance to treatment and screening facility on mammography utilization will be more pronounced and better understood as larger areas are studied. An alternate approach to individual route calculations is being developed to process the large numbers of mammography records more efficiently. Service areas will be computed from treatment and screening centers, and mammography events will be assigned driving distances and times from GIS overlay functions. This approach should be possible after analysis of the current routing results establishes distance classifications that can be used to size service areas.

Additional improvements could be made to the GIS methods in the study in both geocoding and routing. Further address cleanup and comparison with MVD records would increase the geocoding match rate. Better protocols are needed to handle non-standard addresses such as rural routes and post office boxes. Assigning average travel speeds to a better road network would allow more realistic route choices and the ability to determine driving times instead of just driving distances. Additional GIS studies comparing distance to actual facility used with distance to closest facility would refine our measures of geographic variation in mammography utilization.

Acknowledgments

This research was supported by the National Cancer Institute SEER Special Study N01-PC-67007 and Breast Cancer Surveillance Consortium Project U01-CA-69976.

References

1. Rosenberg RD, Lando JF, Hunt WC, Darling RR, Williamson MR, Linver MN, Gilliland FD, Key CR. 1996. The New Mexico Mammography Project: Screening mammography performance in Albuquerque, New Mexico, 1991 to 1993. *Cancer* 78(8):1731-9.
2. Fortney J, Rost K, Warren J. *Comparing alternative methods of measuring geographic access to health services*. Working paper.