EnviroInfo 2000: Umweltinformatik '00 Umweltinformation für Planung, Politik und Öffentlichkeit

Regionale Untersuchungen im Gesundheitsbereich mit Hilfe von CARESS

M. Rohde[1], J. Kieschke[1], I. Wellmann[1] und F. Wietek[1]

Abstract

The possibilities of CARESS (CARLOS Epidemiological and Statistical Data Exploration System), implementing support for the statistical and epidemiological data analysis, are presented in two examples. CARLOS is the corresponding project, started in 1993, to install a population-based cancer registry in Lower-Saxony. Special emphasis is put on the support of spatial data analysis which is the main focus of the system. A separate geographical data server provides both high-quality spatial data as well as facilities to manipulate these data which stem from the field of geographical information systems. The first example is a regionally based prognosis for cancer incidence in the administrative districts of Lower-Saxony, the second is an exploration for the influence of stables on the cronic disease of paranasal sinus.

1. Einleitung

Mit der Ansammlung immer größerer Datenbestände ist in letzter Zeit insbesondere in der betriebswirtschaftlichen Anwendung die systematische und komfortable Datenanalyse in das allgemeine Interesse gerückt. Daten aus vielen operativen Datenbanken werden in Data Warehouses integriert. Sogenannte OLAP- (Online Analytical Processing) Tools dienen auf der Basis der Data Warehouses zur Entscheidungsunterstützung (Chaudhuri/Dayal 1997). Diese Werkzeuge modellieren den betrachteten Datenbestand i.a. als eine Menge multidimensionaler Datenräume (auch Datenwürfel oder Data Cubes), die durch eine Reihe kategorieller Attribute (Dimensionen), wie z. B. Produktgruppe, Verkaufsgebiet oder -zeitraum, aufgespannt werden und zu den jeweiligen Kombinationen von Kategorien auf verschieden granularen Aggregierungsebenen (wie Landkreis, Gemeinde oder Bundesland) die betreffenden Maßzahlen (summarische Attribute), wie Umsatz, Gewinn, Verkaufszahlen etc., enthalten. Im Vordergrund bei der Nutzung von OLAP-Tools steht eine möglichst einfache Bedienbarkeit durch "Nicht-Statistiker" oder "Nicht-Programmierer", de-

[1] Registerstelle des Epidemiologischen Krebsregisters Niedersachsen, OFFIS, Escherweg 2, 26121 Oldenburg, Internet: http://www.krebsregister-niedersachsen.de,
email: Rhode | Kieschke | Wellmann | Wietek@krebsregister-niedersachsen.de

nen auf der Suche nach interessierenden Informationen eine flexible Navigation durch den Datenbestand zu ermöglichen ist.

Nun ist die Problematik der Analyse mehrdimensionaler Datenräume sicher nicht auf die Auswertung von Unternehmensdatenbanken beschränkt. So stellen etwa in der Epidemiologie Neuerkrankungszahlen oder Sterblichkeitsraten nach Alter, Geschlecht, Untersuchungsregion, betrachtetem Zeitraum und Art der Erkrankung typische auszuwertende multidimensionale Datenräume dar. Wie in der Betriebswirtschaft, dem originären Anwendungsgebiet von OLAP, ist auch hier Fachkräften aus verschiedenen Disziplinen, wie Sozialwissenschaften, Medizin, Dokumentation etc., eine komfortable, explorative Analyse des Datenmaterials zu ermöglichen.

In diesem Beitrag wird gezeigt, wie das epidemiologische Informationssystem CARESS (CARLOS Epidemiological and Statistical Data Exploration System) (Wietek 1999a) als OLAP-Tool mit direktem Zugriff auf relevante Geo-Daten vielfältige Unterstützung für geostatistische Auswertungen bietet. Weiterhin wird gezeigt, wie CARESS auf dieser Grundlage für verschiedene regionale Untersuchungen im Gesundheitsbereich genutzt werden kann.

Als Beispiele sollen eine durchgeführte regionale Prognose für die Krebserkrankungszahlen in den Landkreisen Niedersachsens und eine geplante Untersuchung zum Einfluss von Massentierhaltungsanlagen auf die Entstehung chronischer Nasennebenhöhlenentzündungen beschrieben werden. Mit Hilfe von CARESS können durch solche Untersuchungen Versorgungsstrukturen geplant sowie Entscheidungen zu Maßnahmen gegen gesundheitsschädigende Emissionsquellen (z.B. durch eine vorausschauende Bauleitplanung) getroffen werden.

2. Datenanalyse mit CARESS

Das Informationssytem CARESS ist von OFFIS (Oldenburger Forschungs- und Entwicklungsinstitut für Informatik-Werkzeuge und -Systeme) im Rahmen des Aufbaus des Epidemiologischen Krebsregisters Niedersachen entwickelt worden, um die statistisch-epidemiologische Analyse eines Krebsregister-Datenbestands zu unterstützen. Es wird bereits in den Krebsregistern der Bundesländer Niedersachsen, Hamburg und Schleswig-Holstein eingesetzt, wobei das System von verschiedenen Anwendergruppen sowohl für Ad-Hoc-Anfragen und die explorative Analyse des Datenbestands wie auch für Qualitätskontrolle, Inzidenz- bzw. Mortalitäts-Monitoring und Gesundheitsberichterstattung verwandt wird.

Ein Hauptanwendungsgebiet der Krebsregister und damit auch ein Schwerpunkt von CARESS liegt auf der raumbezogenen Datenanalyse. Mit der Analyse von räumlichen und zeitlichen Verteilungen von Erkrankungsdaten erhofft man sich vor allem, potentielle Gesundheitsrisiken zu entdecken. Das Basismodell für die interaktive Datenanalyse, dass CARESS zugrunde liegt, ist in Abbildung 1 zu sehen. Durch das Modell wird ein iterativer Analysevorgang unterstützt. Die Visualisierung

der Ergebnisse einer Untersuchung kann zu einer Interpretation oder zu neuen Analyseschritten führen, wobei die bereits vorgenommenen Selektions- und Analyseschritte erneut verwendet werden können. CARESS kapselt den Zugriff und die Integration der verschiedenen Arten der Daten: die als Einzelfälle vorliegenden Falldaten und zusätzliche Informationen wie Umwelt-, Geo- und demographische Daten. Im Analyseverlauf werden komplexe statistische Daten mit geographischen Daten kombiniert.

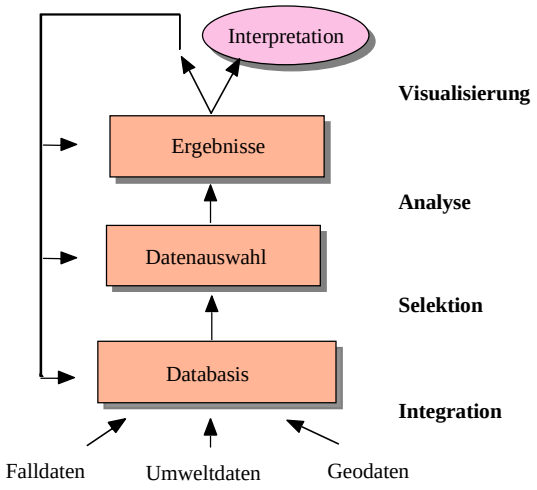

Abbildung 1: Interaktive Datenanalyse

Das Konzept von CARESS basiert auf einem multidimensionales Datenmodell (Wietek 1999b). Insbesondere ist CARESS ein OLAP-Tool, dass auf aggregierten Daten (z.B. Fallzahlen nach Regionen, Alter und Geschlecht) in Form von Data Cubes arbeitet. Mit Hilfe von Dimensionshierarchien auf den klassifizierenden Attributen (Region, Zeit, Alter, Geschlecht, Diagnose und anderen Attributen) wird die flexible Aggregierung auf verschiedenen Granularitätsstufen ermöglicht.

CARESS ist nicht auf den Einsatz in Krebsregistern beschränkt, sondern kann allgemein im Rahmen ökologischer Studien der deskriptiven Epidemiologie, in denen von der Betrachtung des Individuums abstrahiert wird, eingesetzt werden.

3. Verwaltung geographischer Daten in CARESS

Die interaktive Arbeit mit einer großen Anzahl komplexer geographischer Objekte erfordert es, Konzepte geographischer Informationssysteme zu integrieren. Spezielle Datenhaltungskonzepte für die große Menge an Daten mit im wesentlichen lesenden

Zugriff sind wichtig für die geforderte Effizienz bei der typischen Suche nach Mustern in der explorativen Datenanalyse. CARESS benötigt nicht nur die Bereitstellung der Geodaten, sondern auch verschiedene Manipulationsmöglichkeiten wie die Gruppierung und die Bestimmung benachbarter Regionen, den Test, ob ein Punkt in einem Polygon liegt, die Berechnung des Abstands zwischen Regionen und Möglichkeit, geographische Objekte mit beliebigen Attributen zu versehen. Solche Operationen bilden die Grundlage, die betrachtete Studienpopulation in Gruppen einzuteilen (z.B. entlang eines Flusses oder um ein Kraftwerk herum) und Erkrankungsraten in Beziehung mit Daten der geographischen Objekte zu setzen und so potentielle Risikofaktoren zu untersuchen (Abbildung 2). Neben den Operationen ist die Qualität und Vollständigkeit der geographischen Daten wichtig für die räumliche Datenanalyse. Die wichtigste Datenquelle für CARESS bilden ATKIS-Daten (AdV 1994), die neben administrativen Grenzen auch eine große Anzahl detaillierter geographischer Objekte der Deutschen Grundkarte anbieten.

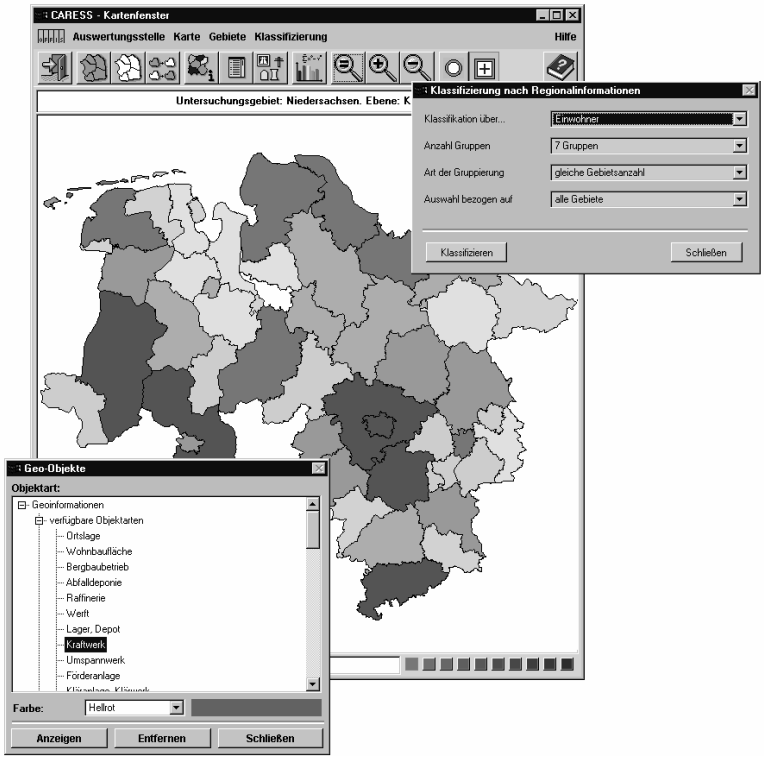

Abbildung 2: Geographische Daten in CARESS

Das geographische Informationssystem InterGIS (Friebe 1997) verwaltet die geographischen Daten für CARESS und bildet die Schnittstelle zu den ATKIS-Daten. InterGIS ist in seiner Architektur auf die Verwaltung sehr großer Datenbestände und lesenden Zugriff optimiert. In CARESS werden in einer korrespondierenden Geo-Bibliothek die geographischen Operatoren und in der Geo-Dimensionshierarchie eine Menge von vordefinierten Standardgruppierungen für Gebiete bereitgestellt.

4. Räumliche Statistik in CARESS

Entsprechend den Empfehlungen des Symposiums „Methoden regionalisierter Beschreibung und Analyse von Krebsregisterdaten“ (Baumgardt-Elms et al 1997) sind in CARESS verschiedene Maßzahlen und Verfahren zur Beschreibung von räumlichen oder räumlich-zeitlichen Häufungen von Krebserkrankungen implementiert worden. Bei den Verfahren der sogenannten globalen Clusteranalyse wird die Homogenität der Inzidenz im gesamten Untersuchungsgebiet untersucht, bei der lokalen Clusteranalyse werden einzelne Regionen der Clusterung identifiziert. Die Clusterindizes Moran’s I (Moran 1948), Moran’s I_{pop} (Oden 1995), Geary’s c (Geary 1954), die Statistik von Ohno & Aoki (Ohno et al. 1979) und der D-Test (Möhner 1991) sind Indizes der globalen Clusteranalyse. Sie beschreiben die räumliche Korrelation, ob also benachbarte Regionen ähnliche Ausprägungen einer Maßzahl aufweisen. Der Index Moran’s I berechnet zum Beispiel die Kovarianz von beliebigen Maßzahlen benachbarter Regionen im Verhältnis zur Gesamtvarianz:

$$I = N \frac{\sum_{i=1}^{N} \sum_{i=1}^{N} w_{ij} (x_i - x)(x_i - x)}{w \sum_{i=1}^{N} (x_i - x)^2}$$

(N ist die Anzahl der Regionen, x_i die Maßzahl für die Region i, x das arithmetische Mittel der Maßzahlen über alle Gebiete. w_{ij} ist ein Element der Adjazenzmatrix (w_{ij}=1, wenn die Regionen i und j benachbart, sonst w_{ij}=0) und w die Anzahl von benachbarten Regionen).

Die Statistik von Potthoff-Wittinghill (Potthoff/Wittinghill 1966) betrachtet dagegen die Heterogenität von Krebsraten allgemein, ohne die Nachbarschaft von Regionen zu berücksichtigen. Der Test von Knox (Knox 1964) beschreibt, ob Fälle, die räumlich benachbart sind, auch zeitlich nahe beieinander liegen. Dieser letzte Test ist speziell für die Betrachtung von Infektionskrankheiten wichtig. Der Algorithmus von Besag & Newell (Besag Newell 1991) ist ein Algorithmus der lokalen Clusteranalyse und hilft, Cluster zu lokalisieren. Der Algorithmus kann folgendermaßen beschrieben werden:

- Auswahl der Clustergröße $C=md$, wobei d die durchschnittliche Fallzahl pro Gebiet sei und m eine positive ganze Zahl (C ist ein Vielfaches der durchschnittlichen Fallzahl).
- Clustertest für jede einzelne Region r_i:
  - Die Regionen $r_1,...r_N$ werden nach der Entfernung von r_i geordnet, $r_{(1)}$ bezeichne dabei r_i, $r_{(2)}$ die Region, deren Mittelpunkt die geringste Entfernung vom Mittelpunkt von $r_{(1)}$ hat, $r_{(3)}$ bezeichne die zweitnächste Region, ...
  - Bestimmung der Zahl M, so dass die Summe der Fallzahlen der Gebiete $r_{(1)}$, $r_{(2)}$, ..., $r_{(M)} \geq C$ ist. Ein kleiner Wert von M deutet – bei entsprechend geringen Bevölkerungszahlen – auf einen Cluster um $r_{(1)}$ hin.
  - Signifikanztest unter Berücksichtigung der Bevölkerung und unter der Annahme einer Poisson-Verteilung der Fälle und einer erwarteten Anzahl von Fällen, die sich aus der Rate des gesamten Studiengebiets errechnet.

Weitere Verfahren zur lokalen Clusteranalyse und Verfahren zum Glätten von thematischen Karten werden in Zukunft implementiert.

5. Beispiel 1: Prognostizierte Krebserkrankungszahlen

5.1 Problemstellung

Da Krebserkrankungen häufiger im fortgeschrittenen Lebensalter entstehen, ist selbst unter der Annahme eines gleichbleibenden alterstandardisierten Krebserkrankungsrisikos in den nächsten Jahren ein deutlicher Anstieg der absoluten Erkrankungszahlen zu erwarten. Aufgrund regionaler Unterschiede in der Altersstruktur des Bevölkerungsaufbaus ist entsprechend mit einem regional unterschiedlichen Anstieg der Krebserkrankungszahlen zu rechnen. Informationen hierüber sind wichtig für die Planung von Versorgungsstrukturen.

5.2 Methode

Zur Berechnung der erwarteten Fallzahlen in den Landkreisen Niedersachsens für die Jahre 1993 bis 2011 wurden die altersspezifischen Inzidenzraten des Krebsregisters des Saarlandes (für die Jahre 1991 bis 1995 zusammengefasst) zugrunde gelegt und angenommen, sie würden über den untersuchten Zeitraum konstant bleiben. Außerdem wurde angenommen, dass es innerhalb Niedersachsens keine regionalen Unterschiede in dem Erkrankungsrisiko gäbe (Der Effekt regionaler Unterschiede im Erkrankungsrisiko ist für die Entwicklung der Fallzahlen vernachlässigbar klein).

Dann ergibt sich für die erwartete Fallzahl e_i der Region i in einem Jahr und bei einem Geschlecht g:

$$e_i = \frac{\sum_{a \in A} R_a n_{i,a}}{\sum_{a \in A} n_{i,a}}$$

(A ist hierbei die Menge der betrachteten Altersgruppen, R_a die Rohe Rate des Saarlands in der Altersgruppe a bei Geschlecht g und $n_{i,a}$ die Bevölkerungszahl in der Region i, der Altersgruppe a, bei Geschlecht g und in dem entsprechenden Jahr). Die Annahmen für die Bevölkerungsentwicklung wurden der regionalen Vorausschätzung der Bevölkerung Niedersachsens des Niedersächsischen Landesamtes für Statistik (NLS) für die Jahre 1993 bis 2011 entnommen (NLS 1999), in der neben der Basisbevölkerung vom 1.1.1993 auch die für die 8. koordinierte Bevölkerungsvorausberechnung - Variante 2 - angesetzten Annahmen über die Zuwanderung eingeflossen sind.

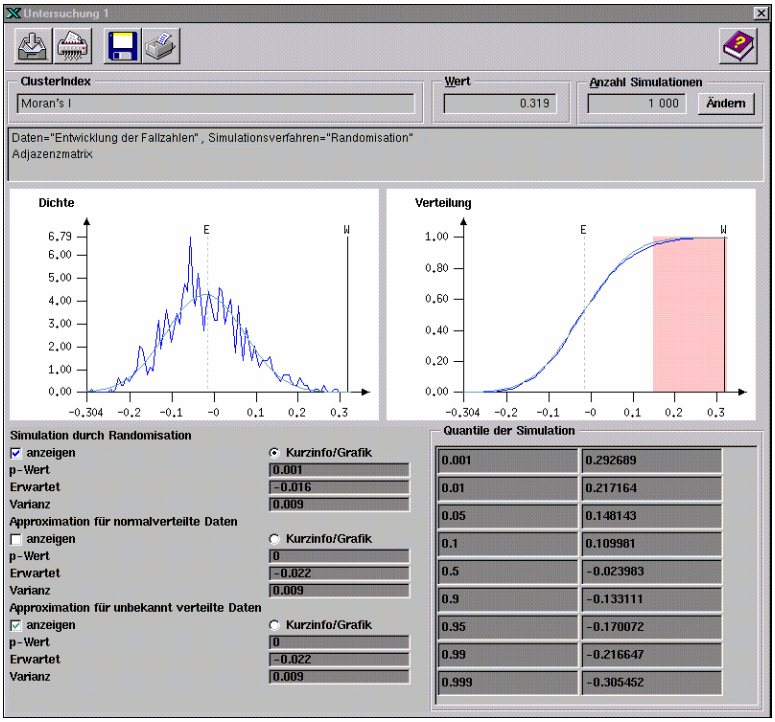

Abbildung 3: Clusterindex Moran's I in CARESS

Zur Klärung, ob es räumliche Häufungen in der Änderung der Krebsneuerkrankungen geben wird, wurde als Teststatistik der Clusterindex Moran´s I (s.o.) berechnet (Abbildung 3). Als Maßzahl x_i wurde die aufgrund der Bevölkerungsentwicklung prognostizierte prozentuale Veränderung der Krebserkrankungszahl verwandt und davon ausgegangen, dass sie die Realisation einer Zufallsvariablen X_i ist. Informationen über die Verteilung von $I=I(X_1,..., X_N)$ wurden mit Hilfe von Simulationen (durch 1000 zufällige Permutationen der Originaldaten) gewonnen (Rohde et al. 1999).

5.3 Ergebnisse dieser Untersuchung

Während für Niedersachsen insgesamt zwischen 1993 und 2011 ein Bevölkerungsanstieg von 5,9 % prognostiziert wird, ist mit einem Anstieg der Erkrankungszahlen um 26,1 % zu rechnen, wobei dieser Anstieg auf Kreisebene zwischen 9,1 % (Goslar) und 48,2 % (Cloppenburg) liegt. Abbildung 4 zeigt kartographisch die geschätzte Entwicklung der Fallzahlen auf Kreisebene in Niedersachsen.

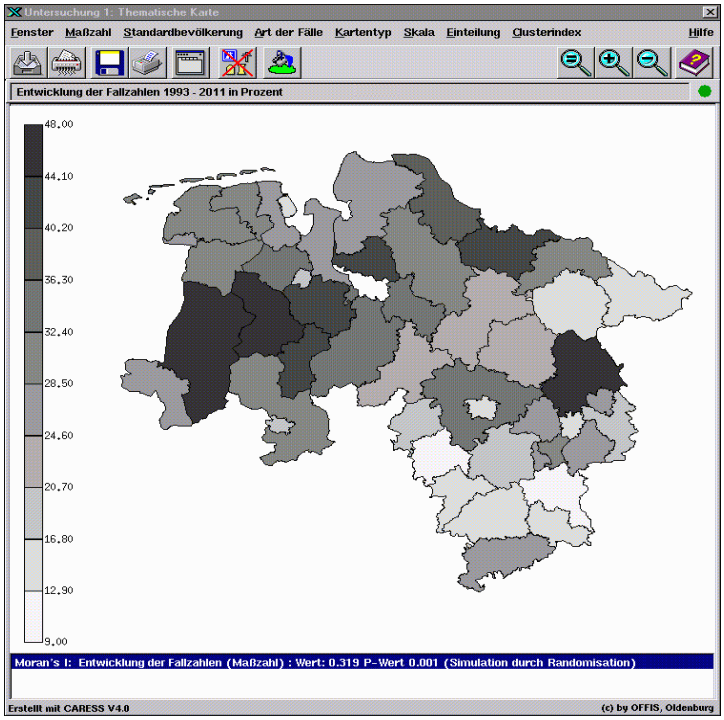

Abbildung 4: Prozentualer Anstieg der Neuerkrankungen (1993-2011)

Es liegt eine räumliche Häufung der Gebiete mit einem besonders hohem erwarteten Anstieg der Erkrankungszahlen vor: Das Ergebnis der Teststatistik gemäss Moran´s I liegt mit 0,319 deutlich jenseits des erwarteten Wertes von -0,016, so dass die Nullhypothese „es gibt keine räumliche Häufung in der (prognostizierten) Zunahme der Krebserkrankungszahlen“ (bei einem vorgegebenen Signifikanz-Niveau von 0,01) abgelehnt wird.

Ein Ausbau der großen Behandlungszentren in den Städten ist nur bedingt anzustreben, da gerade dort die geringste Zunahme zu erwarten ist, stattdessen sollte insbesondere in den ländlichen Regionen darauf geachtet werden, dass eine wohnortnahe Versorgung der Betroffenen möglich ist. Hierbei kann das Netz der in Niedersachsen flächendeckend tätigen „Nachsorgeleitstellen“ helfen.

6. Beispiel 2: Einfluss von Massentierhaltungsanlagen auf Nasennebenhöhlenerkrankungen

6.1 Problemstellung

Da in bestimmten ländlichen Gebieten wie im Oldenburger Raum eine starke Zunahme an Patienten aller Altersklassen mit chronischen und operationspflichtigen Nebenhöhlenerkrankungen beobachtet wird, soll in einer „post-hoc-Untersuchung“ der Einfluß von Massentierhaltungsanlagen auf die Entstehung chronischer Nebenhöhlenentzündungen untersucht werden (durch diese Anlagen werden im großen Umfang Substanzen emittiert, die nachweislich Allergien auslösen können).

6.2 Methode

Um die Zusammenhänge zu analysieren, wird mit Hilfe von CARESS eine regionale Untersuchung der Erkrankungen auf der Basis vorhandener Datenquellen der HNO-Zentren erfolgen.

Nach Ermittlung der landwirtschaftlichen Bebauung in der Region Weser-Ems sowie Modellrechnungen für die Belastung durch Emissionen für die untersuchten Gebiete (Raumbezug) kann mit CARESS die Inzidenz der registrierten Fälle in Abhängigkeit vom Abstand zur Risikoquelle analysiert werden. Da in dieser Untersuchung einzelfallbezogene Daten vorliegen, lassen sich genaue Analysen durchführen, beispielsweise der Vergleich von beobachteten und erwarteten Fällen, gewichtet nach dem Abstand von den Emissions-Punktquellen. Entsprechende Bevölkerungs-Bezugsdaten müssen durch die Einwohnermeldeämter bereitgestellt werden (es werden Angaben für Gebiete, die durch den Abstand von einer Risikoquelle und nicht durch Zugehörigkeit zu einer administrativen Einheit definiert sind, benötigt). In CARESS steht für diese Analyse auf der Basis von Gauss-Krüger-Koordinaten ein

einheitlicher geographischer Raumbezug zur Verfügung, so dass die Wohnort- und Adressangaben bestimmten geographischen Objekten zugeordnet werden können.

Zur Untersuchung der räumlichen Verteilung von Erkrankungsfällen bei Punktquellen stehen außerdem verschiedene statistische Verfahren zur Verfügung. So lassen sich die Verfahren von Stone und von Besag-Newell (s.o.) auch auf Regionen anwenden, deren Grenzen durch konzentrische Kreise um die Punktquelle definiert sind. Falls genauere Informationen zu einem Ausbreitungsmodell vorliegen (z.B. Informationen über häufige Windrichtungen usw.), sind auch andere Modelle für die Aufteilung der Regionen möglich.

7. Zusammenfassung/Diskussion

In diesem Beitrag haben wir gezeigt, welche Möglichkeiten CARESS mit der Integration aller Analyseschritte für regionale Untersuchungen und Planungen im Umwelt- und Gesundheitsbereich bietet. Das System kann ohne großen Aufwand um neue Datenanalyseverfahren erweitert werden. Insbesondere kann CARESS durch die Flexibilität und Erweiterbarkeit auch in anderen verwandten Bereichen außerhalb epidemiologischer Krebsregister eingesetzt werden.

Die Vielzahl der implementierten Clusteranalyse-Verfahren lässt dem Anwender zwar viele Freiheiten, wirft allerdings auch Fragen bezüglich der Nutzung auf (welches Verfahren ist in welcher Situation das geeignetste?), deren Beantwortung eine der zukünftigen Aufgaben sein muss.

Die Beispiele sind nur erste Ansätze für weitergehende Analysen. Geschlechtsdifferenzierte und diagnosespezifische Untersuchungen für die prognostizierten Erkrankungszahlen sollen im Zusammenhang mit Krankenhausstatistiken genauere Grundlagen für die Krankenhausplanung liefern. Mit CARESS gefundene Cluster für Nebenhöhlenerkrankungen können Ausgangspunkt einer Fall-Kohorten-Studie sein, mit der die Bedeutung verschiedener Einflussfaktoren untersucht werden kann.

8. Literatur

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV) (1991): ATKIS-Gesamtdokumentation, Teil D, ATKIS-Objektartenkatalog. Niedersächsisches Landesvermessungsamt, Hannover

Baumgardt-Elms, C., Schümannn, M., von Manikowsky, S., Haartje, U. (eds.) (1997): Symposium "Methoden regionalisierter Beschreibung und Analyse von Krebsregisterdaten", Hamburg, March 1996. Temmen Verlag, Bremen

Besag, J., Newell, J. (1991): The Detection of Clusters in Rare Diseases. Journal of the Royal Statistical Society 154 (1): 143-155

Chaudhuri, S. und Dayal, U. (1997): An overview of data warehousing and OLAP technology. SIGMOD Record, 26(1):65-74, 1997.

Friebe, J. (1997): Eine GeoServer-Architektur zur Bereitstellung geographischer Basisdaten im Internet. In: Dittrich, K.R., Geppert, A. (eds.): Datenbanksysteme in Büro, Technik und Wissenschaft, Ulm, March 1997. Springer Verlag, Berlin, Heidelberg: pp. 251-60

Geary, R.C. (1954): The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician 5: 115-145

Knox, E.G. (1964): The Detection of Space-time Interactions. Applied Statistics 13: 25-29

Möhner, M. (1991): A Global Rank Test for Geographical Clusters of Disease. Biometrical Journal 33: 317-323

Moran, P.A.P. (1948): The Interpretation of Statistical Maps. Journal of the Royal Statistical Society, Ser. B 10: 243-251

Niedersächsisches Landesamt für Statistik (NLS) (1999): Statistik-Datenbank CD 1998.

Oden, N. (1995): Adjusting Moran's I for Population Density. Statistics in Medicine 14: 17-26

Ohno, Y., Aoki, K., Aoki, N. (1979): A Test of Significance for Geographic Clusters of Disease. International Journal of Epidemiology 8: 273-281

Potthoff, R.F., Wittinghill, M. (1966): Testing for Homogeneity. Biometrika 53: 167-190

Rohde, M., Wietek, F., Wellmann, I. Benutzerdokumentation CARESS, Version 4.0. Technischer Bericht OFFIS, Oldenburg, November, 1999.

Wietek, F. (ed.) (1999): Spatial Statistics for Cancer Epidemiology - the Cancer Registry's Epidemiological and Statistical Data Analysis System (CARESS), Universität Bielefeld, Fakultät für Gesundheitswissenschaften 1997. Environmental Health Surveillance. Results of an International Workshop. ecomed-Verlag, Landsberg

Wietek, F. (ed.) (1999): Modelling Multidimensional Data in a Dataflow-Based Visual Data Analysis Environment, Heidelberg 1999. Advanced Information Systems Engineering. Proceedings of the 11th International Conference (CAiSE'99). Springer Verlag, LNCS 1626