Poster

Sächsische Statistiken mit SAS

Bernd Richter
Statistisches Landesamt
des Freistaates Sachsen
Ref.:Soziales

Macherstr. 63
01917 Kamenz
Bernd.Richter@statistik.sachsen.de

Thomas Dittmann
Statistisches Landesamt
des Freistaates Sachsen
Ref.:Bevölkerungsfortschreibung,
Prognose
Macherstr. 63
01917 Kamenz
Thomas.Dittmann @statistik.sachsen.de

Zusammenfassung:

Das Statistische Landesamt des Freistaates Sachsen ist eine obere Landesbehörde im Geschäftsbereich des Sächsischen Staatsministerium des Innern mit derzeit etwa 450 Beschäftigten. Dienstort ist Kamenz, eine Kreisstadt mit knapp 19 Tausend Einwohnern, ca. 45 km nordöstlich von Dresden.

Die Aufgabe des Dienstleisters Statistisches Landesamt besteht darin, auf gesetzlicher Grundlage Daten über Massenerscheinungen zu erheben, zu sammeln, aufzubereiten und zu veröffentlichen. Diese Daten liefern fachlich und regional tief gegliederte Informationen, die für Planungs- und Entscheidungsprozesse benötigt werden. Die Konsumenten der insgesamt etwa 250 Statistiken stammen hauptsächlich aus Politik und Verwaltung, der Wirtschaft und ihren Verbänden sowie aus Wissenschaft und Forschung. Aber auch die Medien und die Bürger nutzen unsere Statistiken. Neben der Bereitstellung von Daten nach einem bundeseinheitlichen Standardkatalog werden individuelle Auswertungen angeboten.

Hier kommt es vor allem auf flexible Auswertungsmöglichkeiten an, die schnell für spezifische Fragestellungen angepasst werden können. Dazu gehören z.B. die Fehlerkontrolle der Ausgangswerte mit PROC MEANS und die Erstellung von statistischen Tabellen mit PROC TABULATE.

Darüber hinaus bestimmen zunehmend wissenschaftliche Analysen, Prognosen und Modellrechnungen unsere Tätigkeit. Zur Bewältigung der Datenmengen und für mathematisch-statistische Analysen wird SAS seit Gründung des Amtes 1992 sowohl auf dem Großrechner als auch als PC-Lösung eingesetzt.

An ausgewählten Beispielen aus dem Bereich der Bevölkerungs- und Sozialstatistiken sollen Anwendungsfelder von SAS verdeutlicht

werden, die diesem Bereich zuzuordnen sind. Es werden Ergebnisse vorgestellt, die folgende Statistiken betreffen:

- Asylbewerberleistungsstatistik,
- Statistik der Empfänger von Hilfe zum laufenden Lebensunterhalt (Sozialhilfeempfängerstatistik),
- Bevölkerungsprognose unter Nutzung der Statistiken über Zu- und Fortzüge, Geburten und Sterbefälle sowie zum Bestand der Bevölkerung.

Ausgangspunkt sind plausibilisierte Einzeldatensätze der verschiedenen Erhebungen. Das sind je nach Erhebung pro Berichtsjahr zwischen 35 und 400 Tausend Datensätze mit je bis zu 50 Merkmalen, wie zum Beispiel Geburtsjahr, Geschlecht, Staatsangehörigkeit und Wohnort.

Keywords: Faktoranalyse, PROC FACTOR, Clusteranalyse, PROC FASTCLUS, Statistik der Asylbewerber, Sozialstatistik, Bevölkerungsstatistik.

1. Statistik der Asylbewerber - Erweiterte Analysemöglichkeiten durch Nutzung zusätzlicher Datenquellen und multivariater statistischer Auswertungsmethoden

Häufig werden im Rahmen der amtlichen Statistik Informationen zu Personen erfasst, welche staatliche Hilfen erhalten (z.B. Sozialhilfeempfänger, Wohngeldempfänger). Für die Zwecke der Planung der benötigten Mittel zur Gewährung dieser Leistungen steht oft die Aufgabe, diese Personengruppen näher zu charakterisieren. Hier sichert die mehrdimensionale Betrachtungsweise die Aufdeckung von Besonderheiten, die durch einfache Tabellierung bzw. Berechnung von eindimensionalen Maßzahlen verborgen bleiben. Im Folgenden wird diese Vorgehensweise für die Statistik der Asylbewerber vorgestellt. Darüber hinaus wird demonstriert, wie durch die Nutzung allgemein zugänglicher Informationen der Aussagegehalt der Ausgangsdaten verbessert werden kann.

Sachsen beherbergt etwa 12 000 Asylbewerber. Der Informationsbedarf besteht hier, neben allgemeinen Aussagen zur Anzahl und zur Entwicklung, in der genaueren Beschreibung der Menschen, welche hier Asylbewerberleistungen erhalten. Die eindimensionale Datenanalyse zeigt, dass Asylbe-

werber in der Regel jung und männlich sind, überwiegend keine Familie haben und meistens aus Ländern kommen, die in denen nicht unbedingt eine allgemeine Gefährdung vermutet wird. Daraus formt sich ein Meinungsbild, welches nicht in jedem Fall der Realität entspricht.

Unter Nutzung von SAS wurde der Aussagegehalt der Ausgangsdaten durch folgende Schritte erweitert:

Nutzung allgemein verfügbarer Informationen

Ein wichtiges Kriterium zur Charakterisierung der sich im Lande befindlichen Asylbewerber ist das Herkunftsland. Steht doch die Frage, inwieweit von den einzelnen schutzsuchenden Personen Deutschland als neues Aufenthaltsland unter geographischem Aspekt, d.h. der Nähe zum Heimatland ausgewählt wurden. Außerdem liefert die Information über das Herkunftsland weitere wichtige Informationen, die sich aus der dort vorhandenen politischen Lage ergeben. Ist doch die dort vorhandene potentielle Gefährdung als ein wesentlicher Grund für die berechtigte Stellung eines Asylbewerberantrages anzusehen.

Das Herkunftsland der Asylbewerber wird durch den Länderschlüssel erfasst, der a priori rein qualitative Aussagemöglichkeiten bietet. Dieser dreistellige Länderschlüssel ist logisch aufgebaut. Den europäischen Staaten sind die Schlüsselnummern 100 bis 199 zugeordnet, die afrikanischen Staaten belegen 200 bis 299, die asiatischen 400 bis 499. Die Kontinente Amerika bzw. Australien und Ozeanien spielen für unsere Auswertungen keine Rolle. Damit kann man den Länderschlüssel im weitesten Sinn als Maß für die Entfernung von Deutschland interpretieren und er kann in die Auswertungen als kategoriales Merkmal aufgenommen werden.

Als weiteres kategoriales Merkmal wurde die Stufe der Gefährdung im Herkunftsland aufgenommen. Dabei wurden folgenden Kriterien angewendet:
0 = keine Gefahr, 1 = Kriege, Unruhen in der Vergangenheit (z.B. Bangladesh), 2 = akute Gefährdung im Moment (z.B. Afghanistan). Grundlage für die Bewertung waren die allgemein zugänglichen Informationen.

Aufdeckung versteckter Zusammenhänge durch die Faktoranalyse

Die Faktoranalyse erfolgte auf der Basis der Rangkorrelationskoeffizienten nach Spearman. Ansonsten wurden die Berechnungen im Wesentlichen mit den Standardeinstellungen von SAS für PROC FACTOR vorgenommen. Es wurden die Merkmale ermittelt, die in etwa denselben Aussagegehalt haben.

Auf dieser Grundlage erfolgt eine Merkmalsreduktion für die weiteren Analysen.

Gruppenbildung – Herausfilterung typischer Personengruppen durch Anwendung der Clusteranalyse

Die Clusteranalyse erfolgte unter Nutzung der Prozedur PROC FASTCLUS. Die maximale Clusteranzahl wurde auf zehn festgesetzt. Die Eingabewerte wurden standardisiert.

Im Ergebnis der Clusteranalyse konnten ca. die Hälfte der Asylbewerber bestimmten Personengruppen zugeordnet werden:

Herkunftsland Europa – drei Gruppen
Alle weisen eine erhöhte Gefahrenstufe auf und unterscheiden sich durch das Alter.

Herkunftsland Afrika – zwei Gruppen
Eine Gruppe ist durch eine erhöhte Gefährdung und vorhandenes Vermögen gekennzeichnet, die andere Gruppe weist kein Vermögen auf und kommt aus Ländern ohne Gefährdung.

Herkunftsland Asien – zwei Gruppen
Kinder aus Gebieten mit höchster Gefahr, die Mitglied einer Großfamilie sind bilden eine typische Gruppe. Eine weitere Gruppe bilden Personen aus nicht gefährdeten Gebieten, in der Regel ohne Kinder.

Die weiteren Personen, die nicht zugeordnet werden konnten, kommen fast alle aus Asien. Für weitere Ergebnisse siehe Tabelle1.

Tabelle 1: Spezifizierung typischer Personengruppen von Asylbwerbern in Sachsen

Cluster-nummer	Herkunfts-region	Charakterisierung der Personen-gruppe	Personen	Anteil in Prozent
1	Europa	erhöhte Gefährdung, Haushaltsvorstand, ca. 1960 geb.	564	5
2	Europa	ca. 1979 geb.	887	7
3	Europa	viele Mitglieder im Haushalt, ca. 1980 geboren	1 971	16
5	Afrika	Vermögen vorhanden, erhöhte Gefährdung	445	4
8	Afrika	Kein Vermögen, keine Gefährdung	867	7
10	Asien	Kinder aus Großfamilien, erhöhte Gefährdung	568	5
4	Asien	Pers. aus nicht gefährdeten Gebieten, ohne Kinder	552	4
6	Asien	nicht eindeutig	1 831	15
7	Asien	nicht eindeutig	2 559	21
9	Asien	nicht eindeutig	2 166	17
	Insgesamt		**12 410**	**100**

2 Bestimmung von Aussagekomponenten zur Charakterisierung der sozialen Lage in Sachsen

Die Darstellung des Bereiches der sozialen Sicherung erfolgt im Rahmen der amtlichen Statistik durch eine Reihe von Erhebungen, welche die wichtigsten Gebiete der sozialen Sicherung (z.B. Sozialhilfe, Wohngeld) und

verwandte Gebiete (Jugendhilfe, Kranken- und Pflegeversicherungen) umfassen.

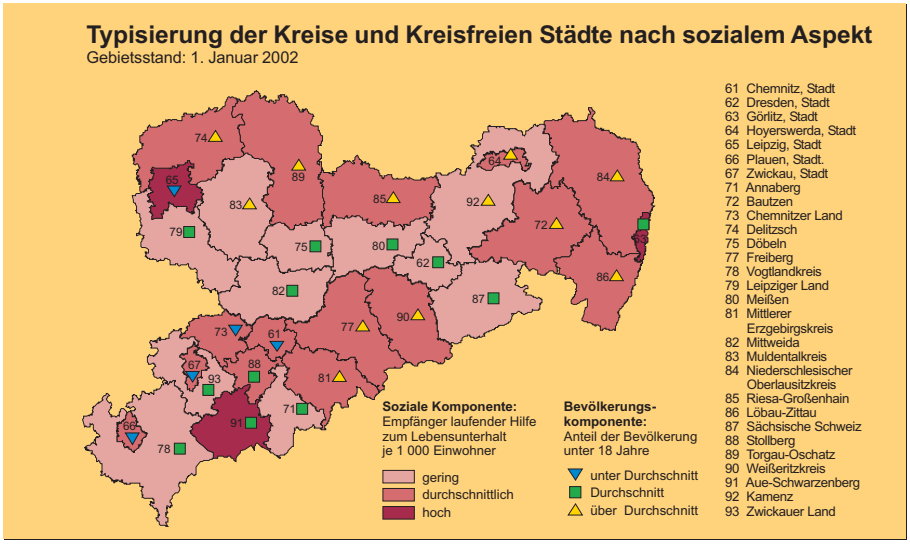
Die Nachfragen zu diesen Statistiken haben in den letzten Jahren zugenommen. Das heißt, die steigenden Anforderungen an die Verwaltung zur Lösung der im sozialen Bereich auftretenden Probleme spiegeln sich in einem erhöhten Informationsbedarf wider.

Im Rahmen der Einzelerhebungen werden sehr viele Merkmale erfasst. Dies führt dazu, dass bestimmte Merkmale zwar eigenständige Aussagewerte haben, in ihrer Aussage als soziale Indikatoren aber fast identisch sind. Zum Beispiel weisen Gemeinden mit einer hohen Anzahl von Empfängern von laufender Hilfe zum Lebensunterhalt auch eine große Anzahl von Wohngeldempfängern auf. Solche Beziehungen sind nicht immer so offensichtlich. Mit Hilfe mathematischer Methoden können versteckte Korrelationen aufgedeckt werden und es kann der Aussagegehalt der verfügbaren Informationen zum sozialen Bereich genauer definiert werden.

Im vorliegenden Fall wurde versucht, die Aussagedimensionen der verschiedenen Kennziffern zum sozialen Bereich zu ermitteln. Dabei wurde eine Faktoranalyse durchgeführt.

Es lagen Ausgangsdaten auf der Ebene der Kreise bzw. kreisfreien Städte vor. Da die meisten Indikatoren altersspezifisch sind, wurde die Alterstruktur der zu betrachtenden territorialen Einheiten mit in die Untersuchung einbezogen.

Die Faktoranalyse zeigte, dass eine sinnvolle Klärung der Zusammenhänge schon mit zwei Faktoren erreicht wird: Dabei enthält der erste Faktor alle Indikatoren für die soziale Lage, der zweite die Strukturmerkmale zur Bevölkerung. Auf der Grundlage der so spezifizierten Faktoren erfolgte eine Typisierung der territorialen Einheiten. Die Ergebnisse sind in Übersicht 1 dargestellt.



Daraus folgt, dass Analysen zur sozialen Lage auf der Basis weniger Kernmerkmale erfolgen müssen. Die Einbeziehung mehrerer Merkmale führt nur zur Überbetonung von bestehenden Unterschieden.

3 Gebietstypisierung für regionalisierte Bevölkerungsprognosen durch Anwendung der Clusteranalyse

Im Bereich Bevölkerungsstatistik werden seit 1996 regelmäßig regionalisierte Bevölkerungsprognosen gerechnet. Diese sind Planungs- und Entscheidungshilfen in verschiedensten Bereichen, in denen Bevölkerung als Bedarfsträger oder Nutzer von Infrastruktur auftritt.

Der Vorteil dieser Prognosen ist, dass hier flächendeckend für ganz Sachsen regional gegliederte Prognosedaten bereitgestellt werden. Die Berechnungen erfolgen nach einem einheitlichen Prinzip, werden aber dennoch unter Berücksichtigung der regionalspezifischen Eigenschaften ermittelt.

Bei der Regionalisierung ist es aus Gründen der statistischen Zuverlässigkeit jedoch notwendig, die kleinsten regionalen Einheiten (hier 530 kreisangehörigen Gemeinde) zu größeren Aggregaten zusammenzufassen. Ziel ist es, in sich homogene, untereinander jedoch heterogene Aggregate mit einer für

Prognosen ausreichend großen Basisbevölkerung (mind. 50 000 Einwohner) zu schaffen.

Die für eine solche Typisierung der Gemeinden notwendigen Indikatoren wurden aus den nur im Statistischen Landesamt so umfassend vorliegenden Daten ermittelt. Berücksichtigung fanden das Geburtenverhalten, die Sterblichkeit, die Wanderungsverflechtungen sowie die Alters- und Sozialstruktur der Bevölkerung, ausgewählte Arbeitsmarktdaten und Veränderungen im Wohnungsbestand.

Die benötigten Daten betreffen verschiedene Statistikbereiche und liegen im Statistischen Landesamt in unterschiedlichen Datenbankformaten vor. Mit SAS lassen sich diese unterschiedlichen Daten-Formate einlesen und über den eindeutigen Gemeindeschlüssel zu einer einheitlichen Datenbasis für alle Gemeinden in Sachsen zusammenführen. Damit steht eine Datenbasis bereit, die die unterschiedlichen Statistiken vereint und für weitere Auswertungen und Analysen in der Faktor- und Clusteranalyse zur Verfügung steht.

Auch hier wurden mittels der Faktoranalyse in der Datenbasis redundante Variablen gefunden durch deren Ausschluss aus der Analyse keine qualitativen Verluste entstehen, aber die Übersichtlichkeit verbessert wird. Anhand der Merkmale dieser reduzierten Datenmasse werden die Gemeinden entsprechend ihren Verhaltensparametern in Gruppen (Cluster) eingeteilt.

Die Gruppierung der Gemeinden erfolgte mit Hilfe der Clusteranalyse unter Berücksichtigung der o. g. Merkmale der Bevölkerung in den einzelnen Gemeinden. Die Clusteranalyse für die 530 Gemeinden wurde mit dem Programm SAS durchgeführt. In einem mehrstufigen Prozess kam dabei die Prozedur PROC FASTCLUS zum Einsatz. Um die unterschiedlichen Daten verwenden zu können, wurden sie standardisiert.

Abfiltern der Gemeinden, die Ausreißer darstellen

Ausreißer, also Gemeinden mit extremen Verhaltensmustern in der Bevölkerung, lassen sich durch die Clusteranalyse sehr schnell finden. Sie werden Cluster zugeordnet, die wenig Elemente haben, einen großen Radius (SAS-Ausgabevariable _RADIUS_) und einen großen Abstand zum nächsten Cluster (SAS-Ausgabevariable _GAP_) ausweisen. Die Idee besteht nun darin, vorerst möglichst viele Cluster zuzulassen und damit die Gemeinden, die eine große Entfernung zu den Gemeinden mit ausgewogenen Parametern (Kernbereich der Gemeinden) aufweisen, in Cluster mit wenigen oder sogar nur einer Gemeinde zusammenzufassen. Das Kerngebiet der Gemeinden wird jedoch nur in wenige Cluster aufgeteilt. Wir haben in der ersten Stufe

mit PROC FASTCLUS und der Option MAXCLUSTERS eine Clusteranzahl von 50 zugelassen.

Die Ermittlung der Cluster mit potentiellen Ausreißern erfolgt mit Hilfe eines GAP-Radius-Scatterplots, bei dem für jedes Cluster der _RADIUS_ und der _GAP_ in Abhängigkeit von der Anzahl der Elemente abgebildet werden. Es muss nun festgelegt werden, ab welcher Mächtigkeit (=Anzahl der Gemeinden je Cluster) ein Cluster kein Ausreißercluster mehr ist und damit in die Clusteranalyse des Kerngebietes einfließen kann. In unserem Fall wurden alle Gemeinden der Cluster, die weniger als 9 Gemeinden enthielten als Ausreißer behandelt. Insgesamt wurden 29 Gemeinden mit extremen Verhaltensparametern gefunden und von der weiteren Analyse vorerst ausgeschlossen.

Jedes Cluster wird durch einen so genannten SEED gekennzeichnet. Die PROC FASTCLUS lässt als Option zu, bei Übergabe einer Datei mit SEEDs nur die Gemeinden in die Clusteranalyse einzubeziehen, deren SEEDs übermittelt werden. Damit ist es möglich, allein durch die Angabe der SEEDs alle diejenigen Gemeinden auszublenden, die zu einem der Ausreißercluster gehören. Über diese Auswahl können die Daten sofort in die folgende 2. Stufe überführt.

Clustern des Kerngebietes mit geringer Clusterzahl

Es muss festgestellt werden, wie viele Cluster maximal gebildet werden sollen und wie groß der STRICT-Wert eines solchen Clusters sein sollte, um möglichst optimale, d.h. in sich homogene und untereinander heterogen Cluster zu erhalten. Der STRICT-Wert legt dabei eine Abstandsgrenze der Beobachtungen zum Clustermittelpunkt (SEED) fest, so dass bei überschreiten des Wertes die Beobachtung dem Cluster nicht zugeordnet wird, zeigt aber als Alternative den nahesten Cluster an.

Der STRICT-Wert sollte nah an den _GAP_- und _RADIUS_-Werten der größeren Cluster liegen, die Anzahl der Cluster sollte in diesem Fall 15 nicht übersteigen, da sonst der Aufwand für die Gestaltung der vielfältigen Prognoseannahmen zu groß (steigt z. T. quadratisch zur Clusteranzahl) wird.

Damit ergibt sich eine Optimierungsaufgabe. Zur Lösung steht ein von SAS geliefertes Optimalitätskriterium, das Pseudo-F-Statistik-Kriterium zur Verfügung, welches maximiert werden soll.

Mit Hilfe eines SAS-Makros wurden die Clusteranalysen für die vorgegebenen Parameterbereiche von bis zu 15 Clustern und einem gewissen Intervall von STRICT-Werten generiert. Dann wurde anhand des für die Pseudo-F-

Statistik erreichten Maximums über alle gerechneten Clusteranalysen der optimale Parametersatz (Clusteranzahl, STRICT-Wert) bestimmt. Im Ergebnis der Optimierung ergaben sich als optimale Zuordnung 6 Cluster für die Gemeinden im Kernbereich.

Nach dem Festlegen der Gebietstypen (Cluster) des Kernbereiches wurden die Ausreißer analysiert. Dazu flossen zusätzlich weitere, nichtnumerische Informationen über die spezifischen Eigenschaften dieser Gemeinden in die Entscheidungsfindung ein. Die „Ausreißer"-Gemeinden wurden daraufhin in 3 Gruppen aufgeteilt, die sich durch hohe Neubautätigkeit (1 Gruppe) und Gemeinden mit Asylbewerber- und Aussiedlerheimen (2 Gruppen) unterscheiden.

Die insgesamt 16 Cluster (9 Cluster aus der Clusteranalyse und die 7 Kreisfreien Städte des Freistaates Sachsen) bildeten die Basiseinheiten für die Erarbeitung der Annahmen zur voraussichtlichen Entwicklung der demografischen Verhaltensparameter. Das bedeutet, dass die einzelne Gemeinde zwar nicht mit ihren „eigenen" aber dennoch mit hinreichend ähnlichen Werten in die Berechnung eingeht. Durch die Auswertung der Prognoseergebnisse nach Regionen wird der daraus resultierende Typisierungsfehler aber erfahrungsgemäß gut ausgeglichen.

Die anschließende Aufbereitung und Auswertung der Prognoseergebnisse (2,5 Mio. Datensätze) erfolgt ebenfalls mit SAS. Auf der einen Seite steht eine Abfragemaske im SAS zur Verfügung, über die jeder Mitarbeiter im Fachbereich mit Mausklick in einer Windowsumgebung Ergebnisse der Prognose selektieren, verdichten und zur weiteren Verarbeitung exportieren kann. Auf der anderen Seite wird die programmiertechnische Umsetzung über SAS-Connect realisiert, wodurch die gegenüber dem PC höhere Rechengeschwindigkeit des Großrechners genutzt wird.

Diese Anwendungsmöglichkeiten bieten auch uns im Statistischen Landesamt des Freistaates Sachsen eine flexible und effiziente Nutzung der Ressourcen und unterstützen nachhaltig die Verbreiterung unserer Angebotspalette bei der Bereitstellung von standardisierten und nutzerspezifischen Datenanforderungen für alle Interessierten aus politischen, wissenschaftlichen und privaten Bereichen.