

Von Daten mit geografischen Punktzuordnungen zu Kartendarstellungen - ein (fast) universelles Makro

Thomas G. Grobe
AQUA-Institut
Maschmühlenweg 8-10
37073 Göttingen
Thomas.Grobe@Aqua-Institut.de

Zusammenfassung

Häufig werden Daten mit einer Georeferenzierung in sehr unterschiedlichem Umfang einzelnen geografischen Punkten mit einer zugleich variablen regionalen Dichte zugeordnet. Selbst bei umfangreichen Daten ist zumeist eine Zusammenfassung von Daten bzw. von bestimmten Regionen erforderlich, um Ergebnisse sinnvoll geografisch darstellen zu können. Geeignetes Kartenmaterial ist nicht immer verfügbar.

Dem Autor standen aggregierte Daten zu etwa zehn Prozent der deutschen Bevölkerung mit regionaler Zuordnung zu mehr als 8.000 Postleitzahlen (PLZ) inklusive geografischer Längen- und Breitenangaben zur Verfügung. Ziel der Auswertungen waren u.a. differenzierte bildliche Darstellungen von geschlechts- und altersadjustierten regionalen Erkrankungshäufigkeiten. Hierfür wurde ein umfängliches SAS-Makro entwickelt, welches nach geeigneter Anpassung auch ohne die Verfügbarkeit von Kartenmaterial weitgehend universell für andere Daten und Regionen eingesetzt werden kann.

Startpunkt des Makros sind Daten, bei denen jedem durch eine Längen- und Breitenangabe definierten Punkt jeweils weitgehend beliebig viele Erhebungsmerkmale zugeordnet sind (z.B. Geo-Koordinaten zu einzelnen PLZ mit Angaben zur regionalen Populationsgröße sowie zur Zahl von Erkrankten, jeweils differenziert nach Geschlechts- und Altersgruppen, inklusive einer Bundeslandzuordnung).

Das vorgestellte Makro bestimmt zunächst eine rechteckige Fläche, welche alle Punktkoordinaten mit verfügbaren Daten einschließt und scannt diese Fläche anschließend schrittweise in einer gewählten Auflösung (z.B. in einem 5 km-Raster). Im Umkreis jedes Rasterpunktes werden dabei Angaben zu ursprünglichen Datenpunkten zusammengefasst, bis ein gewähltes Kriterium erfüllt ist. Anschließend werden die zusammengefassten Angaben zu einzelnen Rasterpunkten zur Berechnung von regionalen Kennwerten verwendet (z.B. zur Berechnung geschlechts- und altersstandardisierter Erkrankungsraten).

Abschließend werden die Kennwerte zur Erstellung einer Bilddatei genutzt, wobei Flächen mit unzureichenden Beobachtungszahlen ausgespart bleiben und optional Flächenbegrenzungen (z.B. die Bundeslandgrenzen) sowie Längen- und Breitengrade dargestellt werden können. Damit entstehen Bilddateien mit Kartendarstellungen zu weitgehend beliebigen Regionen in angepasstem Maßstab ohne Rückgriff auf vorbestehendes Kartenmaterial. Probleme sind bei Nutzung des beschriebenen Makros bei Darstellungen zu Regionen in Pol-Nähe sowie bei Regionen unter Einbeziehung der Datumsgrenze zu erwarten, die ggf. durch geeignete Modifikationen gelöst werden müssen.

Schlüsselwörter: SAS-Makro, Kartendarstellung, georeferenzierte Daten

1 Hintergrund

Daten mit geografischen Zuordnungen (Georeferenzierungen) werden in den letzten Jahren mit zunehmender Häufigkeit genutzt.¹ Typischerweise ist bei entsprechenden Daten sowohl der Umfang als auch die regionale Dichte bzw. Verteilung der Informationen recht heterogen. Selbst bei großen Datenbeständen ist für eine Darstellung von regionalen Ergebnissen regelmäßig eine Clusterbildung/Zusammenfassung von zunächst regional differenzierter zugeordneten Daten für bestimmte Regionen erforderlich, um stabile und nicht vorrangig zufallsbedingte Variationen der Ergebnisse darstellen zu können. In anderen Regionen kann im Rahmen derselben Analyse demgegenüber die regionale Differenzierung möglicherweise ausschließlich durch die Auflösung der verwendeten regionalen Zuordnung selbst limitiert sein. Nicht in jedem Fall ist geeignetes Kartenmaterial vorhanden.

Die Motivation zur Programmierung der hier vorgestellten Makro-Sammlung resultierte aus dem Wunsch nach einer bildlichen Darstellung von populationsbezogen ermittelten regionalen Erkrankungsdaten (sowie vergleichbaren Parameter), wobei die regionale Darstellung – abhängig von Umfang und Art der jeweils betrachteten Daten – regional möglichst differenziert erfolgen sollte. Zudem sollten die dargestellten regionalen Kennwerte unabhängig von möglichen Besonderheiten einer regionalen Geschlechts- und Altersstruktur sein. Regionale, zunächst differenzierter zugeordnete Daten sollten daher ggf. derart zusammengefasst werden, dass anschließend eine Adjustierung – hier in Form einer indirekten Standardisierung – durchgeführt werden konnte. Ausreichend differenziertes Kartenmaterial im SAS-Format stand dem Autor nicht zur Verfügung.

Bei Verfügbarkeit von Kartenmaterial im SAS-Format wäre alternativ zum vorliegenden Makro eine Zusammenfassung benachbarter Regionen nach bestimmten Kriterien zu erwägen (vgl. z.B. Babcock 2010). Hingewiesen sei an dieser Stelle auch auf frei verfügbare Programme mit komplexen statistischen Analysemöglichkeiten von geografisch (und zeitlich) referenzierten Daten (z.B. SaTScan, FleXScan), die allerdings i.d.R. zunächst keine bildlichen Darstellungen liefern. Adäquate Verfahren zur statistischen Absicherung regionaler Unterschiede sind in der vorliegenden Makro-Sammlung nicht implementiert – vorrangiges Ziel war die Darstellung von Ergebnissen in Karten.

2 Material und Methoden

2.1 Verwendete Daten

Dem Autor standen für Auswertungen zu regionalen Variationen der Nutzung diagnostischer Maßnahmen sowie von Erkrankungshäufigkeiten aggregierte Populations-, Inanspruchnahme- sowie Diagnoseangaben zu etwa zehn Prozent der deutschen Bevölkerung mit einer regionalen Zuordnung zu mehr als 8.000 Postleitzahlen (PLZ) zur Verfügung. Angaben zu regionalen Populationsgrößen sowie zu Erkrankungshäufigkeiten und

¹ Eine umfangreiche Übersicht zur Nutzung und Analysemöglichkeiten im epidemiologischen Kontext liefert z.B. Auchincloss et.al 2012.

Inanspruchnahmen medizinischer Leistungen lagen mit regionaler Zuordnung differenziert für jeweils 40 Geschlechts- und Altersgruppen vor. Den PLZ konnten geografische Längen- und Breiten-Angaben im Sinne von Zentroiden bzw. Flächenschwerpunkten der PLZ-Zustellregionen zugeordnet werden.²

2.2 Makro – Struktur der Ausgangsdaten

Startpunkt der vorgestellten Makrosammlung bilden allgemein Datentabellen, in denen jedem durch eine geografische Längen- und Breiten-Angabe definierten Punkt in genau einer Tabellenzeile jeweils weitgehend beliebig viele Erhebungsmerkmale zugeordnet sind (z.B. Geo-Koordinaten zu einzelnen PLZ, ergänzende Bundesland- und ggf. auch Kreiszuordnungen mit Namensbezeichnungen sowie geschlechts- und altersspezifische Angaben zu regionalen Populationen und Erkrankungshäufigkeiten). Georeferenzierungen müssen für die Bearbeitung obligat in Dezimal-Schreibweise vorliegen, also als einfache Grad-Angaben mit Nachkommastellen.³

Bei einer Zusammenfassung werden im Makro-Ablauf ggf. regional ursprünglich unterschiedlich zugeordnete Werte numerischer Variablen ggf. summiert (z.B. Angaben zur Population sowie zu Erkrankten), um erst nach dieser regionalen Zusammenfassung adäquate Kennwerte (wie z.B. geschlechts- und altersstandardisierte Raten) zu berechnen. Andere Formen der Zusammenfassung sind aktuell nicht implementiert, was bereits bei einer Vorbereitung von Daten zu bedenken ist. Zu jedem Rasterpunkt (s.u.) können in den im Makro-Ablauf weiter aufbereiteten Daten zudem jeweils nur die nächstliegenden alphanumerischen Variablenwerte gespeichert werden (z.B. die nächstliegende PLZ, Bundeslandzuordnung oder Ortsbezeichnung).

2.3 Basismakro – Ablauf und Funktionen

Das Basismakro (>*Geo_BasisXXX.sas*<)⁴ zur Aufarbeitung regionaler Daten bestimmt im ersten Abschnitt B-A (vgl. bereitgestellte Syntax) zunächst – vereinfacht formuliert – eine rechteckige Fläche, welche alle geografischen Punkt-Koordinaten aus den betrachteten Daten mit relevanten Informationsangaben einschließt (im Beispiel für Deutschland eine Fläche von ca. 713 km x 895 km). Anschließend scannt das Makro diese Fläche schrittweise in einer gewünschten Auflösung (z.B. in einem 5 km-Raster; vgl. Abschnitt B-B). Jedem Rasterpunkt (im Beispiel $143 \cdot 179 = 25.597$ Punkten) werden dabei verfügbare Angaben zu umliegenden, ursprünglichen Punktkoordinaten zugeordnet, bis ein gewähltes Kriterium erfüllt ist (z.B. eine angegebene Populationsgröße erreicht ist), oder bis eine vorgegebene maximale Distanz für eine Zusammenfassung überschritten wird (z.B. ein 40 km Abstand zum Rasterpunkt)⁵. Daten zu ursprünglichen

² Quelle: Deutsche Post Direkt GmbH; www.postdirekt.de/online-services

³ Universell verwendbare geografische Längen- und Breitenangaben erfordern ein geodätisches Referenzsystem. Seit 1989 wird international das WGS 84 (World Geodetic System 1984) verwendet, welches auch Grundlage der Angaben von üblichen GPS-Empfängern und Navigationsgeräten bildet.

⁴ XXX steht hier für die Versionsnummer des Makros, z.B. für 003.

⁵ Distanzen werden gemäß der Haversine Formel berechnet. Alternativ könnte hier in neueren SAS-Versionen zwischenzeitlich auch die SAS-Funktion GEODIST genutzt werden.

Punktkoordinaten innerhalb einer dem Rasterpunkt exklusiv zugeordneten Fläche werden dabei grundsätzlich einbezogen, um alle regional zugeordneten Daten mindestens bei der Ermittlung von Kennwerten zu einem Rasterpunkt zu berücksichtigen. Anschließend werden die Informationen, wie bereits weiter oben beschrieben, zu einer Beobachtung je Rasterpunkt zusammengefasst.

Ein letzter Abschnitt B-C des Basismakros erstellt eine Datei (hier als Shape-Datei bezeichnet), die im Rahmen der späteren Bilderstellung zur Optimierung der (äußeren) Konturen der Kartendarstellung verwendet werden kann.

Im Ablauf des Basismakros werden folgende SAS-Dateien, bei entsprechender Spezifikation auch als permanente Datendateien, erstellt, wobei die Makro-Variabel `&OUTFILE.` eine auswertungsspezifische Benennung der Dateien ermöglicht:

- `&OUTFILE.P` Datei mit Koordinaten und Rastergrenzen zu *allen* Rasterpunkten (je Rasterpunkt eine Beobachtung).
- `&OUTFILE.R` Datei mit zusammengefassten Ergebnissen je Rasterpunkt (*sofern Ergebnisse zum jeweiligen Punkt verfügbar*).
- `&OUTFILE.XY` Übergeordnete Angaben zur Gesamtregion (eine Datenzeile, z.B. Angaben zur Zahl der Rasterpunkte sowie zur geografischen Ausdehnung der Region).
- `&OUTFILE.S` (optionale) Shape-Datei zur Optimierung der Gebietsumrisse.
- `&OPTFILE.` optional permanente (optimierte) Original-Datei mit Index.

Als weitere, optional auch permanente SAS-Dateien wird im Rahmen des späteren Ablaufs die folgenden Datei erstellt:

- `&OUTFILE.RWZS` Datei mit Angaben zur empirischen Verteilung der Kennwerte (wird lediglich zum automatisierten Einfärben vorläufiger Kartendarstellungen verwendet).

2.4 Makro – Spezifikationen

Sowohl Spezifikationen für das Basismakro als auch für alle übrigen Programmabläufe sind in der SAS-Syntax-Datei `>Geo_AufrufXXX.sas<` vorzunehmen. Eine umfangreichere Beschreibung der Spezifikationen ist der genannten Syntax-Datei zu entnehmen. Alle Spezifikationen für den Programmablauf werden bei einer Ausführung der Syntax (z.B. nach Druck auf die Funktionstaste `>F8<`) in Makrovariablen geschrieben. Anschließend lassen sich die insgesamt vier Teilmakros bei Bedarf ohne Angabe weiterer Spezifikationen ausführen.

Ein erheblicher Teil der Spezifikationen dient der Festlegung von Umgebungsvariablen (z.B. zur Angabe von Verzeichnissen für verwendete Originaldaten, Ergebnisausgaben und Programmdateien) sowie der Angabe von Datei- und Variablennamen (z.B. der Benennung der Variablen mit Geo-Koordinaten in den Originaldaten).

Als wesentliche Spezifikationen zur Steuerung des Makroablaufs können u.a. folgende Parameter bzw. Makrovariablen-Werte festgelegt werden:

- **STEP_UNIT** Einheit für Distanzen (>m< oder >km< für Meter oder Kilometer).
- **STEP_DIST** Schrittweite für Rasterpunkte (in der zuvor spez. Einheit).
- **DIST_MAX** maximale Distanz der ursprünglichen Datenzuordnung zum Rasterpunkt bei einer ggf. notwendigen regionale Zusammenfassung von Ergebnissen.
- **CRIT_MIN** minimaler Wert eines Kriteriums, das, ggf. nach Zusammenfassung, zur Ermittlung und Darstellung von Ergebnissen zu einem Rasterpunkt erreicht sein muss (z.B. eine minimale Populationszahl oder Zahl der Erkrankten).

Die aufgeführten Werte bestimmen sehr maßgeblich den Ablauf des Basismakros. Eine kleine Schrittweite bei Daten zu einer geografisch sehr großen Region kann zu extrem langen Berechnungszeiten führen. Erste Programmdurchläufe zu Testzwecken sollten typischerweise mit relativ großen Schrittweiten erfolgen, die zu einer Nutzung von nur wenigen tausend Rasterpunkten führen.

2.5 Makro – Standardisierung

Die den Rasterpunkten nach Durchlauf des Basismakros zugeordneten Angaben können anschließend zur Berechnung von regionalen Kennwerten verwendet werden. Hierzu ist ein (Teil-)Makro in die SAS-Syntax-Datei *Geo_AufrufXXX.sas* integriert, welches über den Aufruf `>%MSTAND;<` gestartet werden kann. Das Makro erlaubt die Berechnung indirekt geschlechts- und altersstandardisierter (Erkrankungs-)Raten bzw. Ratios mit einer Ermittlung von Konfidenzintervallen. Dabei werden als Ergebnisse z.B. relative Abweichungen der regional beobachteten von den regional strukturabhängig erwarteten Erkrankungszahlen ermittelt. Ein Wert von 1 (bzw. 100%) zeigt an, dass die regional beobachteten Werte den nach überregionalen Ergebnissen erwarteten Zahlen entsprechen. Kleinere Werte deuten auf unterdurchschnittliche Ergebnisse hin, größere Werte auf erhöhte Werte. Das Makro hat typischerweise eine sehr kurze Laufzeit und erzeugt eine ausschließlich temporär verfügbare Ergebnisdatei im SAS-WORK-Verzeichnis, die anschließend für Kartendarstellungen genutzt werden kann. Ein Aufruf dieses (Teil-)Makros ist daher obligat vor einer Kartenerstellung erforderlich.

Das Makro erfordert die Spezifikation *mindestens* einer Zähler- und *mindestens* einer Nenner-Variablen, um populationsbezogene Ergebnisse im Sinne von Raten berechnen zu können (z.B. Erkrankungsfälle je Einwohner). In diesem Fall werden rohe, nicht-standardisierte Raten berechnet. Erst bei Spezifikation von mehreren Zähler-Variablen sowie einer identischen Zahl von entsprechenden Nenner-Variablen sind die Ergebnisse entsprechend dieser Differenzierung indirekt standardisiert.

Sollen regional erhobene Werte ohne einen Nenner-Bezug verarbeitet werden (z.B. einfache physikalische Messwerte mit einer Georeferenzierung), kann das (Teil-)Makro gleichfalls verwendet werden, sofern in den Daten (als „Nenner-Ersatz“) eine Variable ergänzt ist, die zu allen Datensätzen mit gültigen Messwerten konstant den Wert 1 enthält. Bei einer regionalen Zusammenfassung werden damit einfach Mittelwerte berech-

net. Zur späteren Kartendarstellung werden dann relative Abweichungen der regionalen Mittelwerte vom überregional ermittelten Mittelwert über alle Messpunkte verwendet. Bei der zuvor beschriebenen Zweckentfremdung des Makros zur indirekten Standardisierung zur Berechnung einfacher Mittelwerte liefern die im Makro implementierten Berechnungen von Konfidenzintervallen selbstverständlich keine sinnvoll interpretierbaren statistischen Kennwerte. Auch bei einer „regulären“ Nutzung ist allerdings zu beachten, dass die errechneten Konfidenzintervalle weder eine Adjustierung für eine ggf. auch multiple Nutzung von ursprünglichen Ergebnissen für Kennwerte zu mehreren Rasterpunkten noch eine Adjustierung für multiples Testen im allgemeinen Sinne enthalten. Die Werte sollten daher eher als Orientierungshilfe und nicht als Beleg für statistische Signifikanzen gelesen werden.

2.6 Makro – Erste Bilddarstellung

Anschließend können die ermittelten Kennwerte zu Rasterpunkten zur Erstellung von binären Bilddateien mit farblich abgestuften Bildpixeln genutzt werden, wobei Flächen mit unzureichenden Beobachtungszahlen ausgespart bleiben.

Ein erstes und typischerweise schnell durchlaufendes Makro `>%MMAP;<` erzeugt vier Bilddarstellungen mit automatisiert angepasster Einfärbung der jeweiligen Karten, die sich an der empirisch ermittelten Lage und Streuung der Kennwerte orientiert. Es wird eine farbig (von grün bis rot) sowie eine in Graustufen (von hell bis dunkel) eingefärbte Karte erstellt. Zudem werden zwei Karten in Grauabstufungen erstellt, in denen ergänzend Regionen mit 95%- sowie 99%-Konfidenzintervallen von Kennwerten farblich gekennzeichnet sind, die den überregional ermittelten Kennwert nicht enthalten, was als Hinweis auf eine nicht rein zufällig beobachtete Abweichung der regionalen Werte von überregionalen Ergebnissen gedeutet werden kann (unter Beachtung der weiter oben beschriebenen Einschränkungen).

Die Dimensionen dieser Bilddateien entsprechen grundsätzlich dem zur Auswertung verwendeten Raster und sind damit typischerweise sehr klein (im zuvor erwähnten Deutschland-Beispiel resultieren Bilder mit $143 \times 179 = 25.597$ Pixeln). Sollen diese Bilder verwendet werden, empfiehlt sich eine nachträglich Vergrößerung in gängigen Bildprogrammen um einen ganzzahligen Faktor (z.B. 10-fach), wobei unbedingt gewöhnlich verwendete Interpolationen ausbleiben sollten, damit Flächen zu Rasterpunkten nach einer Vergrößerung als einfache, gleichmäßig gefärbte Quadrate im Bild erscheinen.

Die Bilddateien werden vom SAS-Makro in einem TGA-Format erstellt. Die Wahl des Formates resultiert ausschließlich aus vorbestehenden Erfahrungen des Autors mit dem Schreiben von Dateien in diesem Format. Bei Bedarf (wie auch für die Anzeige in einem Internetbrowser, s.u.) müssen die Bilder in gängiger Bildbearbeitungssoftware in andere Formate konvertiert werden.

2.7 Makro – Erweiterte Bilddarstellung, Image Maps

Ergänzend oder alternativ zur einfachen Bilderstellung kann eine erweiterte Bilderstellung durchgeführt werden. Implementiert ist diese in `>Geo_KartenXXX.sas<`. Im Rahmen der SAS-Syntax-Datei `Geo_AufrufXXX.sas` kann der Code über `>%FMAP;<` aufgerufen werden.

Die erweiterte Bilddarstellung erlaubt eine individuelle Spezifikation der Bildeinfärbung (der grundlegend eine modifizierbare logarithmische Skalierung der Kennwerte zugrunde liegt). Sehr viel relevanter sind Optionen zur Darstellung von Gebietsbegrenzungen (z.B. von Bundeslandgrenzen) sowie zur Anzeige von Längen- und Breitengraden. Um diese zusätzlichen Informationen ohne Überdeckung der eigentlichen Kennwerte darstellen zu können, werden die Bilder bei der Ausgabe durch das SAS-Programm um einen wählbaren ganzzahligen Faktor skaliert.

Als weiteres Feature kann im Rahmen des Programmablaufs eine HTML-Datei erzeugt werden, die eine sogenannte „Image Map“ enthält. Damit lassen sich bei einer Betrachtung der zuvor erstellten Karten in einem Internet Browser relativ umfangreiche Informationen zu einzelnen Bildbereichen anzeigen, wenn sich der Maus-Zeiger über der entsprechenden Region befindet (z.B. Kennwerte zu einem Rasterpunkt, Namensbezeichnungen zur Region, Links auf Geo-Koordinaten, die nach Mausklick zu einer separaten Anzeige einer Region in Google Maps führen).

Wesentliche Spezifikationen zur Steuerung der erweiterten Kartendarstellung sind:

- **PIC_SCALE** Skalierungsfaktor für das ausgegebene Bild.
- **LOGBAS** Basis zur Logarithmierung von Kennwerten. Bestimmt die Karteneinfärbung (der angegebene Wert kennzeichnet die relative Abweichung, die zur extremen Einfärbung der Karte erforderlich ist. Z.B. sind bei Angabe des Wertes `>2<` regional mindestens doppelt bzw. höchstens halb so hohe Werte wie überregional extrem eingefärbt, hier dunkelrot bzw. dunkelgrün).
- **BORDER** Breite der Darstellung von Grenzen in Pixeln (muss kleiner als Skalierungsfaktor sein, bei Wert 0 wird keine Grenze angezeigt).
- **BORDER_VAR** Merkmal zur Regionskennzeichnung (z.B. eine Bundeslandkennzeichnung). Weicht dieses Merkmal zwischen zwei benachbarten Rasterpunkten ab, wird eine Grenze gezeichnet.
- **BORDER_S** Breite der Darstellung von Subgrenzen in Pixeln (muss kleiner als Skalierungsfaktor sein, bei Wert 0 wird keine Subgrenze angezeigt).
- **BORDER_VAR_S** Merkmal zur Subregionskennzeichnung (z.B. eine Kreis-kennzeichnung). Weicht dieses Merkmal zwischen zwei benachbarten Rasterpunkten ab, wird auch hier eine Grenze gezeichnet. Wird hier die Angabe `>GRAD_RASTER<` gemacht, werden die geografische Rasterpunktzuordnungen zur Darstellung von Längen- und Breitengraden genutzt.
- **DEGREEPART** Angabe eines Teilers geografischer Grad-Angaben. Bei einem Wert von 1 wird je Gradwechsel ein Längen- und Breitengrad dargestellt, bei einem Wert 10 erfolgt z.B. eine Darstellung je Wechsel in 1/10-Grad-Abstufung.

3 Ergebnisse – Darstellungsbeispiele

Genutzt wurde eine vorausgehende Version der hier beschriebenen Makrosammlung bereits für publizierte Darstellungen zur regionalen Nutzung diagnostischer Prozeduren sowie zu regionalen Diagnoseraten in Deutschland (Grobe et. al. 2011). Modifizierte Auszüge der seinerzeit verwendeten Daten bilden die vorrangige Auswertungsgrundlage für die nachfolgend beschriebenen Beispieldarstellungen.⁶ Dargestellt werden exemplarische Ergebnisse zur bevölkerungsbezogenen Häufigkeit von Hypertonie-Diagnosen (ICD10: I10) sowie zur Diagnose hyperkinetischer Störungen (ICD10: F90).

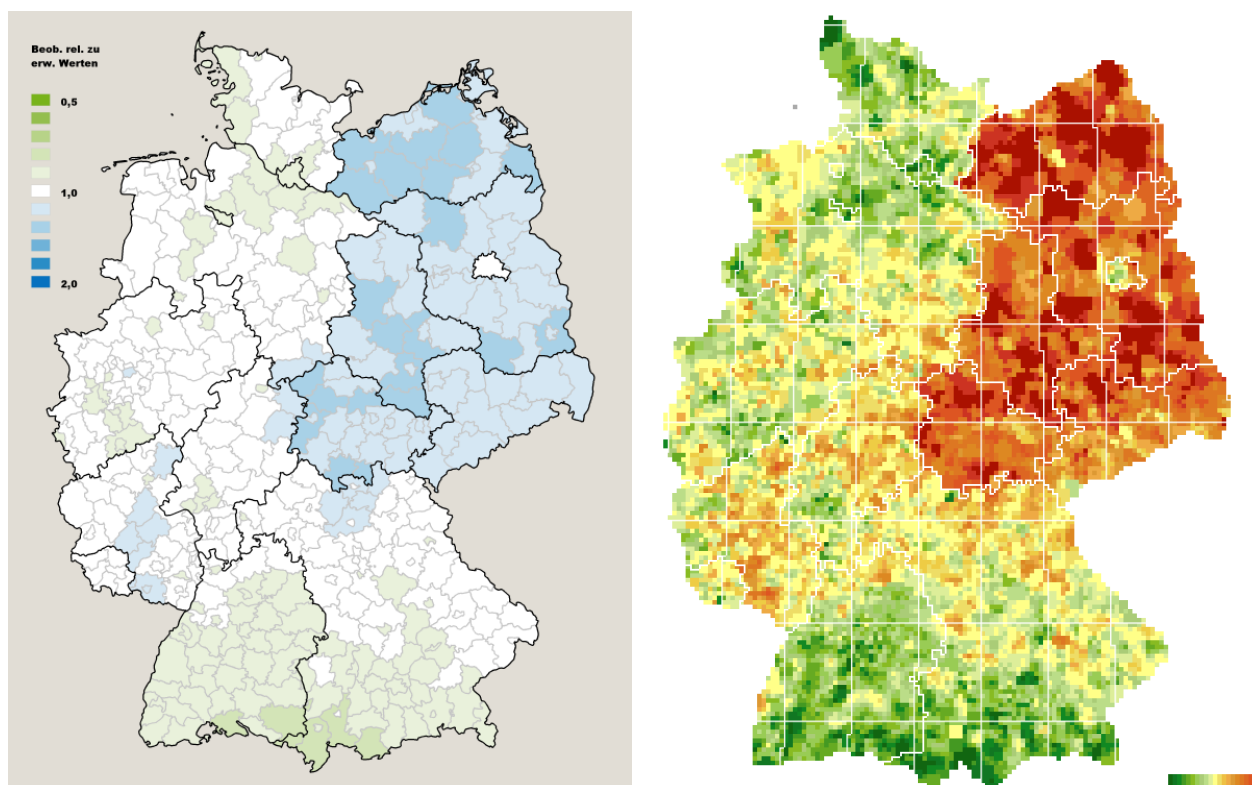


Abbildung 1: Relative Abweichungen regionaler Diagnoseraten von bundesweiten Raten in Kreisen 2012 (SAS Maps) sowie basierend auf Daten mit PLZ-Zuordnung aus dem Jahr 2009 (Makro-Resultat, 5 km-Raster)

Die linke Hälfte der Abbildung 1 zeigt eine Darstellung zu regionalen Abweichungen von Hypertonie-Diagnoseraten in 413 Kreisen Deutschlands (Aufteilung nach Gebietsstand Ende 2008) basierend auf Daten zum Jahr 2012, die vom Autor in SAS/GRAPH unter Rückgriff auf damit verfügbares Kartenmaterial erstellt wurde. Die rechte Hälfte zeigt eine Darstellung basierend auf weitgehend vergleichbaren Daten aus dem Jahr 2009 mit einer Zuordnung zu gut 8.000 5-stelligen PLZ, die mit der hier beschriebenen Makro-Sammlung ohne Rückgriff auf Kartenmaterial erstellt wurde.⁷

⁶ Die Daten mit einer regionalen Zuordnung auf der Ebene von 5-stelligen Postleitzahlen standen dem Autor nur ausnahmsweise und ausschließlich in aggregierter Form zur Verfügung.

⁷ Sofern möglich sollten die Grafiken in der farbigen Version der PDF-Fassung des vorliegenden Beitrags in höherer Auflösung betrachtet werden.

Beiden Abbildungsvarianten zeigen regional indirekt geschlechts- und altersstandardisierte Kennwerte. Die Darstellung der Makro-Variante beruht hier auf Ergebnissen zu einem 5 km-Raster mit den beispielhaft bereits im vorausgehenden Text erwähnten Spezifikationen zugrunde (minimal 1.000 Personen je ausgewiesenem Rasterpunktergebnis, Zusammenfassung von Original-Ergebnissen zu den einzelnen Rasterpunkten im Umkreis von max. 40 km, Bildausgabe um den Faktor 10 auf 1410x1750 Pixel vergrößert, max. Einfärbung ab Überschreitung bundesweiter Ergebnisse um 30 Prozent). Auf einem Notebook benötigte der Durchlauf aller Teil-Makros weniger als 3 Minuten.

Aufgrund der insgesamt relativ hohen regionalen Dichte der geografischen Zuordnungen von PLZ sind die Umrisse Deutschlands auch im Makro-Resultat gut erkennbar wiedergegeben. Feinere Details, wie z.B. die ostfriesischen Inseln, können erwartungsgemäß nicht differenziert werden. Lediglich Helgoland ist als Punkt sichtbar (Ergebnisse werden hier allerdings nicht ausgewiesen, da die Personenzahl unterhalb des gewählten Limits von 1.000 Personen liegt). Die Verläufe der aus den Daten mit PLZ-Zuordnung hergeleiteten Bundeslandgrenzen entsprechen relativ gut den realen Gegebenheiten. Das Grad-Raster bietet eine zusätzliche Orientierung.

Abbildung 2 zeigt Ergebnisse der Makro-Variante wie in Abbildung 1 in größerer Darstellung, für deren Berechnung abweichend zudem ein 2 km-Raster gewählt wurde. Auf einem Notebook benötigte der Durchlauf aller Teil-Makros hier etwa 15 Minuten. Es resultiert ein finales Bild mit 3490x4370 Pixeln. Deutlich ist in diesem Bild erkennbar, dass die geografische Auflösung der Ergebnisse im Falle der häufigen Hypertonie-Diagnosen, die gut ein Viertel der Bevölkerung betreffen, in einigen Regionen maßgeblich auch durch die verfügbare regionale Auflösung der ursprünglichen geografischen Zuordnung der Daten eingeschränkt wird (vgl. insbesondere östliche Bundesländer mit z.T. recht großflächigen PLZ-Gebieten). In anderen Regionen (Ballungszentren mit hoher PLZ- und Bevölkerungsdichte) sind demgegenüber regional sehr differenziert ausgewiesene Ergebnisse erkennbar (vgl. z.B. Berlin). Inhaltlich überraschen bei den Ergebnissen zum Thema Hypertonie die klaren Unterschiede zwischen alten und neuen Bundesländern, was im Rahmen des vorliegenden methodischen Beitrags nicht weiter diskutiert werden soll.

Abbildung 3 stellt regionale Ergebnisse zur Diagnose hyperkinetischer Störungen, besser bekannt unter dem Kürzel ADHS dar (ICD10: F90). Die Diagnose wird in der hier betrachteten Form bei rund ein Prozent der Bevölkerung dokumentiert, ist also erheblich seltener als Hypertonie. Als Kriterium für eine Darstellung von regionalen Ergebnissen wurde eine Mindestzahl von 100 betroffenen Personen vorausgesetzt (nur graduell abweichende Ergebnisse resultieren, sofern eine Bevölkerung von min. 10.000 Personen für eine Darstellung vorausgesetzt wird). Eine maximale regionale Einfärbung erfolgte ab Überschreitung bundesweiter Ergebnisse um 100 Prozent. Inhaltlich auffällig sind bei diesen Ergebnissen insbesondere ausgesprochen hohe Diagnoseraten in einer Region um Würzburg. Die Abbildung verdeutlicht allgemein die Makro-Darstellung bei selteneren Ereignissen. So werden in einigen Randgebieten sowie auch in nördlichen Teilen Brandenburgs keine Ergebnisse ausgewiesen, da das entsprechende Kriterium (hier: ≥ 100 Betroffene im Umkreis von maximal 40 km) nicht erfüllt war (vgl. grau unterlegte Flächen in der Farbdarstellung der Abbildung).

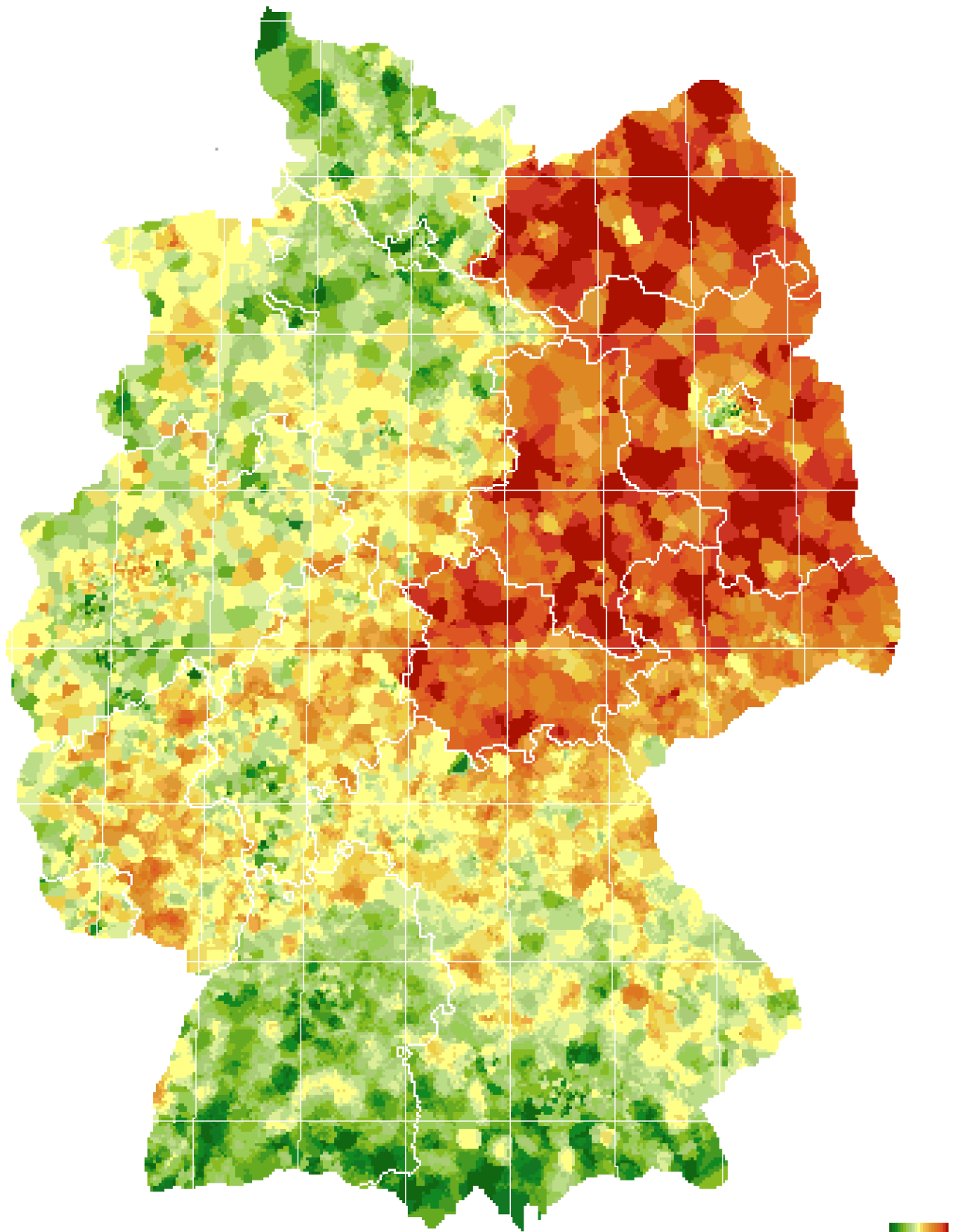


Abbildung 2: Relative Abweichungen regionaler Hypertonie-Diagnoseraten (ICD10: I10) von bundesweiten Raten basierend auf Daten mit PLZ-Zuordnung aus dem Jahr 2009 (Makro-Resultat, 2 km-Raster)

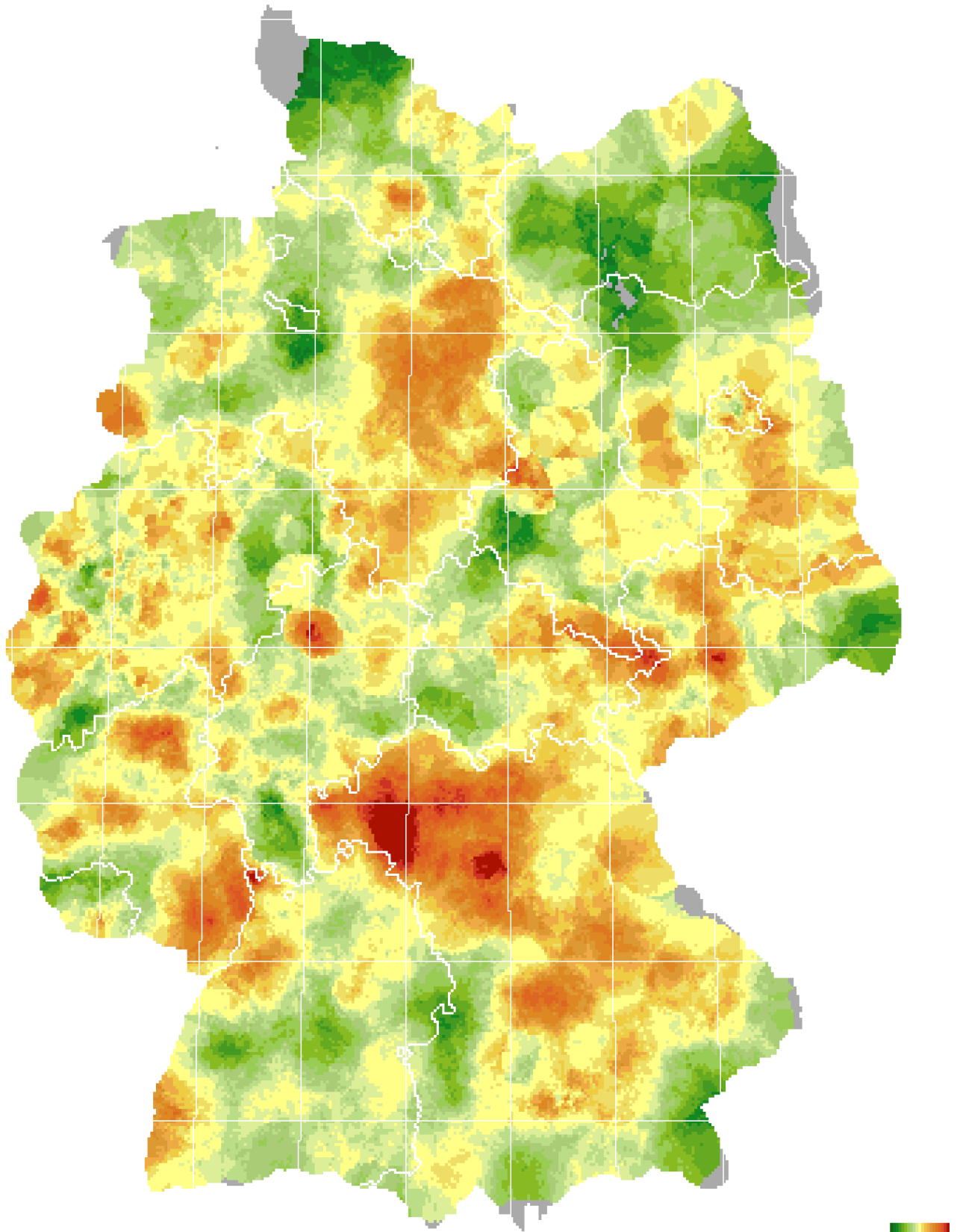


Abbildung 3: Relative Abweichungen der regionalen Diagnoserate hyperkinetischer Störungen (ICD10: F90) von bundesweiten Raten basierend auf Daten mit PLZ-Zuordnung aus dem Jahr 2009 (Makro-Resultat, 2 km-Raster)

Testdatenbeispiele

Für erste Probedurchläufe der hier beschriebenen Makro-Sammlung werden Testdaten bereitgestellt (Datei *plz_b_red_i10.sas7bdat*), die den Test von Funktionen der (Teil-) Makros an realitätsnahen Daten erlauben. Dabei handelt es sich um einen Datenauszug, welcher zu Postleitzahlen mit Lokalisation im Bundesland Berlin für Männer und Frauen in Altersgruppen zwischen 40 bis unter 90 Jahre jeweils Populationsangaben sowie Angaben zur Anzahl der Personen aus der Population mit Diagnose einer Hypertonie (ICD10: I10) enthält. Die Daten werden gemeinsam mit der SAS-Syntax in einer Zip-Datei bereitgestellt. Nach Entpacken dieser Zip-Datei muss die zur Spezifikation des Makro-Ablaufs verwendete Syntax-Datei *>Geo_AufrufXXX.sas<* lokalisiert werden, um dort die Pfadangabe in der Makro-Variable *>STAMV<* anzupassen. Anschließend kann die Datei einmalig mit *>F8<* ausgeführt werden, womit alle erforderlichen Programm-Spezifikationen in Makro-Variablen gelesen werden. Schließlich können die Teil-Makros *%MBAS*, *%MSTAND*, *%MMAP* sowie *%FMAP* durch Markieren und Ausführen nacheinander gestartet werden. Alle Ergebnisse sollten nach erfolgreichen Durchläufen im zunächst leeren Unterverzeichnis *>Ergebnisse003<* auffindbar sein.

Wird ausschließlich die Pfadangabe erfolgreich angepasst, sollten nach dem Ausführen aller Teil-Makros u.a. auch exakt die beiden in der oberen Zeile von Abbildung 4 dargestellten Kartenbilder im Ergebnisverzeichnis verfügbar sein. Längen- und Breitengrade sind in 1/10-Grad Schritten markiert, alle Darstellungen beruhen auf Auswertungen zu Rasterpunkten im Abstand von jeweils 500 m. Die Karte oben links zeigt relative Abweichungen der Hypertonie-Diagnoseraten, wobei regional mindestens Angaben zu 100 Personen berücksichtigt wurden. Die Karte oben rechts deutet durch grüne bzw. rote Einfärbungen sinngemäß an, ob regionale Kennwerte *einschließlich* der beiden Grenzen von 99%-Konfidenzintervallen unterhalb bzw. oberhalb von Bundesland-weiten Durchschnittswerten lagen, was zumindest tendenziell auf nicht-zufällige Abweichungen hindeutet.

Die zweite Zeile der Abbildung 4 enthält vergleichbare Darstellungen, denen, im Gegensatz zur ersten Zeile nicht-standardisierte, „rohe“ Diagnoseraten zugrunde liegen (statt Werten zu einzelnen Geschlechts- und Altersgruppen wurden zusammengefasste Angaben zur Population (Variable: *Ges_pop*) sowie zu Betroffenen (Variable: *Ges_bet*) zur Berechnung verwendet). In einigen Regionen resultieren merklich abweichende Ergebnisse.

Die dritte Zeile zeigt erneut indirekt standardisierte Ergebnisse, wobei je Rasterpunkt Ergebnisse zu mindestens 10.000 Personen berücksichtigt wurden, wobei Angaben im Umkreis von maximal 10.000 m um einen Rasterpunkt berücksichtigt wurden. Durch die Zusammenfassung resultiert eine deutliche Glättung der regionalen Ergebnisse. Aufgrund der hohen Populationszahlen je Rasterpunkt resultieren nach einer entsprechenden Zusammenfassung enge Konfidenzintervalle, womit Konfidenzintervalle ohne Einschluss von Bundesland-weiten Durchschnittswerten relativ häufig erfasst werden.

Der untere Teil in Abbildung 4 zeigt exemplarisch Informationen, die bei einer Betrachtung der Karten in einem Internet Browser für jeden Rasterpunkt, abhängig von der Position des Mauszeigers, angezeigt werden können.

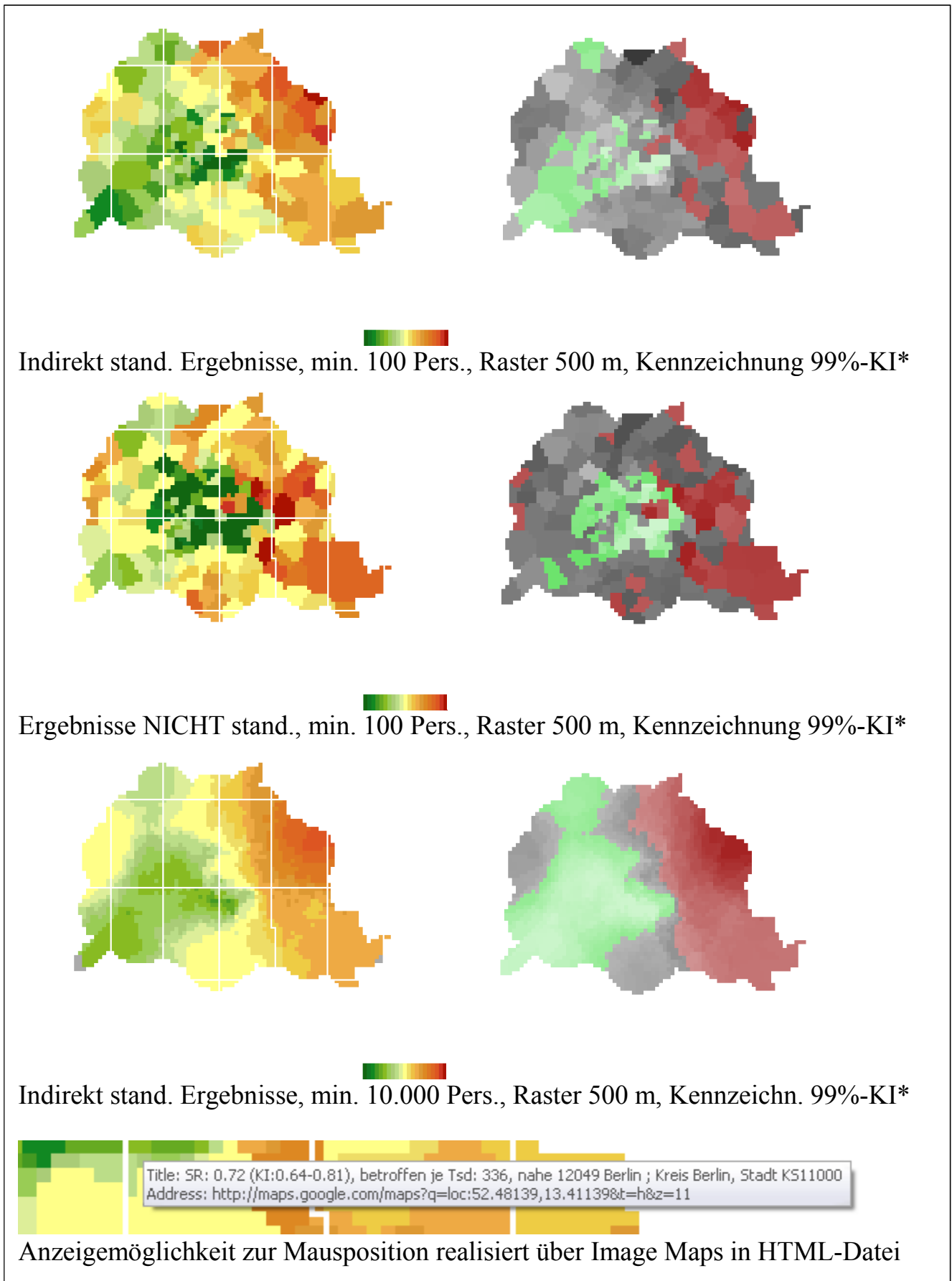


Abbildung 4: Testdatenbeispiele (Hypertonie-Diagnoseraten Berlin, *vgl. Text)

4 Diskussion

Die Makro-Sammlung zur Kartendarstellung wurde im Rahmen des Beitrages vorrangig im Hinblick auf Anwendungsmöglichkeiten und Darstellungsergebnisse vorgestellt. Programm-technisch interessierte Leser finden bei Bedarf weitere Informationen in der bereitgestellten Syntax. Der Ablauf setzt ausschließlich die Installation von SAS/BASE voraus. Das integrierte Modul zur indirekten Standardisierung lässt sich bedarfsweise problemlos auch zur Berechnung „roher“ Raten oder Verarbeitung einfacher Messwerte nutzen. Relativ einfach möglich wäre auch die Integration anderer Adjustierungsverfahren, insbesondere sofern diese auf die ohnehin zu jedem Rasterpunkt bereitgestellten Informationen aufsetzen.

Die unter Rückgriff auf die Makro-Sammlung erzeugten Kartendarstellungen lassen sich vorrangig im Rahmen einer Daten-Exploration und -Visualisierung nutzen. Einschränkungen können durch längere Laufzeiten bei hochaufgelösten Darstellungen resultieren. Eine wesentliche Besonderheit der besprochenen Makro-Sammlung dürfte darin liegen, dass – ausschließlich basierend auf Daten mit Zuordnung zu geografischen Punktkoordinaten – Kartendarstellungen weitgehend universell auch ohne vorgefertigtes Kartenmaterial möglich sind.

Literatur

- [1] A. H. Auchincloss, S. Y. Gebreab, C. Mair, V. A. Diez Roux (2012): A Review of Spatial Methods in Epidemiology, 2000–2010. *Annu Rev Public Health*. Apr 2012; 33: 107–122.
- [2] G. Babcock (2010): Aggregation without aggravation: determining spatial contiguity and joining geographic areas using hashing; SAS Global Forum 2010; Paper 223-2010.
- [3] SaTScan Software: M. Kulldorff and Information Management Services, Inc. SaTScan™ v8.0: Software for the spatial and space-time scan statistics. <http://www.satscan.org/>, 2009.
- [4] FleXScan Software: K. Takahashi, T. Yokoyama, T. Tango. FleXScan v3.1.2: Software for the Flexible Scan Statistic. National Institute of Public Health. Japan, 2013. <http://www.niph.go.jp/soshiki/gijutsu/download/flexscan/>
- [5] T. G. Grobe, H. Dörning, F. W. Schwartz (2011). BARMER GEK Arztreport 2011. Asgard Verlag, St. Augustin, 2011.