Webscraping in der Beherbergungsstatistik — ein Zwischenbericht

Neben Bürgerinnen und Bürgern sind auch Unternehmen für verschiedene amtliche Statistiken auskunftspflichtig. Dies bedeutet für viele Unternehmen zusätzliche Belastungen. Heutzutage können jedoch bereits viele Informationen über Unternehmen und Betriebe durch die Nutzung neuer digitaler Datenquellen wie dem Internet automatisiert erhoben werden. Daten über hessische Unterkünfte mit essentieller Bedeutung für die Beherbergungsstatistik sind dabei häufig auf kommerziellen Online-Buchungsportalen verfügbar und öffentlich zugänglich. Daher wurden im Hessischen Statistischen Landesamt mit Webscraping Daten über hessische Beherbergungsbetriebe aus einem kommerziellen Online-Buchungsportal extrahiert und ausgewertet. Die Chancen des Webscraping von Beherbergungsdaten werden im folgenden Artikel aufgezeigt. ***Von Normen Peters***

Hintergrund

Die Beherbergungsstatistik

Die Beherbergungsstatistik erfasst für alle Beherbergungsbetriebe oberhalb einer Abschneidegrenze (zehn oder mehr Schlafgelegenheiten bzw. Stellplätze) jeden Monat u. a. die Anzahl der Gästeankünfte und Übernachtungen sowie die Herkunft der Gäste (nach Wohnort bzw. Land des Wohnorts). Durch die monatliche Erhebung sowie die Abschneidegrenze nach der Anzahl der Schlafgelegenheiten ist die Pflege des Berichtskreises aufwändig: Betriebsneugründungen und -schließungen sowie die Anzahl der Schlafgelegenheiten müssen verlässlich festgestellt werden, bevor Beherbergungsbetriebe als auskunftspflichtig eingestuft werden können. Die Nutzung neuer digitaler Datenquellen, z. B. mittels Webscraping gewonnen, lassen Potenzial zur vereinfachten Pflege des Berichtskreises erkennen.

Gerade im Tourismusbereich sind eine leichte Auffindbarkeit sowie die Präsenz auf Internetseiten und kommerziellen Buchungsportalen wichtig für das Geschäft der Beherbergungsbetriebe. Eine automatisierte Suche und Auswertung von Internetseiten von Beherbergungsbetrieben – z. B. über verschiedene Buchungsportale – könnte die Aktualisierung des Berichtskreises der Monatserhebung im Tourismus unterstützen.

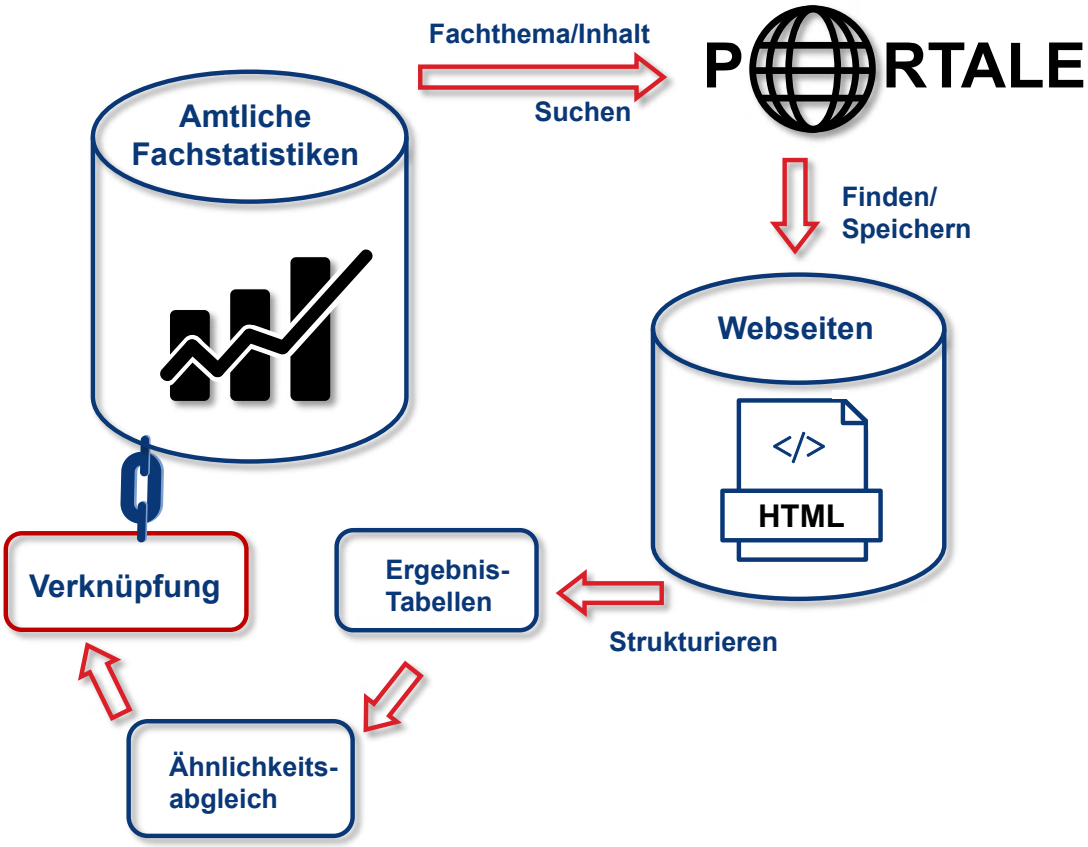
© Modella – Fotolia.com

Webscraping eines kommerziellen Online-Buchungsportals

Das kommerzielle Online-Portal „HRS Holidays – das Ferienhausportal“ ist mittels Webscraping automatisiert nach hessischen Beherbergungsbetrieben durchsucht worden. Das Verfahren baute auf Erfahrungen und Kenntnissen auf, die im Hessischen Statistischen Landesamt (HSL) bereits beim Suchen und Verknüpfen von Unternehmenswebseiten mit den Unternehmen des hessischen Unternehmensregisters erzielt wurden (siehe Hessisches Statistisches Landesamt, 2018).

Die auf den Subwebseiten für die hessischen Beherbergungsbetriebe gefundenen Merkmale wurden strukturiert, gespeichert und ausgewer-

Abbildung 1: Funktionsweise des HSL Online-Portal-Scraping



tet. Kommerzielle Online-Portale eignen sich nicht zuletzt deshalb gut für die treffsichere Informationsextraktion von unternehmensbezogenen Inhalten, weil die beteiligten Betriebe mit dem Betreiber in einer gegenseitigen rechtlichen Beziehung stehen. Da potenzielle Kunden die auf dem kommerziellen Online-Portal gefundenen Produkte und Dienstleistungen rechtsverbindlich bestellen und bezahlen können, ist davon auszugehen, dass die dort angegebenen Daten korrekt sind. Somit waren zahlreiche Prüfalgorithmen, die beim Webscraping von Unternehmenswebseiten die Treffergenauigkeit der gefundenen Internetseiten vor der Verknüpfung mit amtlichen Daten bewerten, nicht erforderlich. Die dem Portal zugrundeliegende Einheitlichkeit der Datenstruktur innerhalb des Portals, verringerte darüber hinaus die Komplexität der Ablaufprogrammierung. Das automatisierte Auslesen des HRS-Portals hatte somit Effizienz-, Kapazitäts- und Geschwindigkeitsvorteile.

Im Folgenden werden die Verfahrensfunktionalitäten sowie die Such- und Extraktionsprozeduren beschrieben. Daran anschließend wird die Anwendung des Webscraping auf ein kommerzielles Online-Portal zum Buchen von Unterkünften in Deutschland näher erläutert und die Ergebnisse werden vorgestellt. Es zeigt sich, dass sehr viele Unterkunftsmerkmale verfügbar sind. Die Auswertung einiger dieser Merkmale folgt in diesem Artikel nicht in erster Linie dem Zweck des Verifizierens oder Falsifizierens einer speziellen Untersuchungshypothese. Es soll vielmehr mit den ausgewählten Merkmalen exemplarisch dargestellt werden, welche Auswertungsmöglichkeiten und welches grundsätzliche Potenzial die Informationsextraktion eines solchen Online-Portals für die Beherbergungsstatistik bietet.

Funktionsweise des Online-Portal-Scraping

Die Inhalte von amtlichen Fachstatistiken, wie Stammdaten, Wirtschaftszweige, Fachinhalte oder Fachthema, können verwendet werden, um in kommerziellen Online-Portalen nach den gewünschten Informationen automatisiert zu suchen. Insbesondere die Art des Fachthemas ist jedoch für die Wahl des passenden Online-Portals ausschlaggebend. In diesem Fall waren die zu suchenden Informationen auf das Thema der hessi-

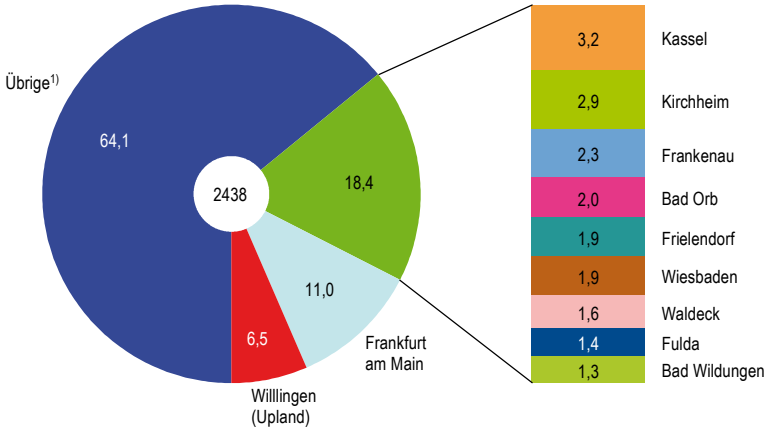
schen Beherbergung und die Wahl des passenden Mediums auf ein kommerzielles Online-Buchungsportal mit möglichst vielen in Hessen befindlichen Unterkünften im Bestand beschränkt. Abbildung 1 illustriert die Funktionsweise der Informationsextraktion von gewünschten Inhalten eines Online-Portals.

Wenn zutreffende Informationen (z. B. Unterkünfte mit Angaben über die Anzahl zur Verfügung stehender Schlafmöglichkeiten in hessischen Gemeinden und Städten) in einem Online-Portal gefunden werden, werden die Subwebseiten gespeichert, auf denen der entsprechende Beherbergungsbetrieb mit seinen Eigenschaften präsentiert wird. Die gespeicherten Subwebseiten des Online-Portals enthalten die gewünschten Daten, jedoch in unstrukturierter Form. Die HTML-Codes werden nun in XML-Dateien umgewandelt. Mithilfe sogenannter CSS- und Xpath-Selektoren oder mit JSON-Applikationen lassen sich alle inhaltlichen Begriffe in den XML-Dateien finden. Diese Selektoren sind durch eineindeutige Zeichenketten benannt und dienen dazu, bestimmte Elemente in HTML/XML-Dateien auszuwählen. Mit zusätzlichen Methoden des Text Mining und der Hilfe „regulärer Ausdrücke“ – eine Art Filterkriterium für Texte – lassen sich die gewonnenen Inhalte in strukturierte Listen oder Tabellen überführen und speichern. Um die amtliche Statistik mit den gespeicherten Inhalten zu ertüchtigen, muss die Verknüpfung über die Stammdaten und ggf. über Ähnlichkeitsabgleiche erfolgen.

Anwendung

Das Scraping von Beherbergungsbetrieben im „HRS Holidays – das Ferienhausportal“ wurde am 21. September 2018 durchgeführt. Nach Auswahl der Region „Hessen“ durchsuchte der Algorithmus in einem ersten Schritt die Hauptseite nach den URLs der Unterwebseiten der hessischen Beherbergungsbetriebe und stellte diese in einer Liste zusammen. In einem zweiten Schritt erfolgte die listenweise Speicherung und weiterführende Analyse der jeweiligen Unterwebseiten, die die Unterkunftsinformationen enthalten. Die so gefundenen Inhalte wurden anschließend strukturiert und in tabellarischer Form gespeichert. Der gesamte Prozess dauerte etwa 20 Minuten.

Abbildung 2: Hessische Beherbergungsbetriebe im HRS-Holiday-Portal nach Städten und Gemeinden (in %)



1) Städte und Gemeinden, die jeweils weniger als 1,3 % der hessischen Beherbergungsbetriebe aufwiesen.

Ergebnisse

Hessische Beherbergungsbetriebe

Mit dem Scraping des Portals „HRS Holidays – das Ferienhausportal“ wurden 2438 hessische Beherbergungsbetriebe gefunden. Zum Vergleich: 6598 Beherbergungsbetriebe wurden für das Jahr 2016 im hessischen statistischen Unternehmensregister geführt. Im statistischen Bericht der hessischen Tourismusstatistik von September 2018 sind 3504 Beherbergungsbetriebe ausgewiesen worden. Folglich enthielt „HRS Holidays – das Ferienhausportal“ nicht alle statistisch in Hessen erfassten Beherbergungsbetriebe. Ein zukünftiger Abgleich mit dem Unternehmensregister sowie das Scraping weiterer Portale wird zeigen, wie viele hessische Beherbergungsbetriebe in Portalen in Hessen präsent sind.

Neben der reinen Zahl der Beherbergungsbetriebe konnten Informationen über die regionale Zugehörigkeit des jeweiligen Betriebes nach „Stadt/Gemeinde“ und „Region“ extrahiert werden. Abbildung 2 zeigt die Verteilung der gefundenen Beherbergungsbetriebe nach hessischen Städten und Gemeinden.

Frankfurt am Main stellt mit 11,0 % im „HRS Holidays – das Ferienhausportal“ den größten Anteil an hessischen Beherbergungsbetrieben, gefolgt von Willingen (Upland) (6,5 %), Kassel (3,2 %), Kirchheim (2,9 %), Frankenau (2,3 %) und Bad Orb (2,0 %). Die Landeshauptstatt Wiesbaden stellt mit 1,9 % einen eher geringen Anteil der hessischen Unterkünfte im HRS-Holiday-Portal. Städte und

Gemeinden, die weniger als 1,3 % der hessischen Unterkünfte zählen, stellen zusammen 64,1 % aller hessischen HRS-Beherbergungsbetriebe.

Verfügbare Merkmale

Jeder Beherbergungsbetrieb konnte in Art und Umfang beliebig viele Unterkunftsmerkmale bezüglich Ausstattung, Gebäudeeigenschaften, Umgebungs- und Freizeitgestaltungshinweise oder Verpflegungsmodalitäten im „HRS Holidays – das Ferienhausportal“ angeben.

Deshalb wiesen die jeweiligen Unterkunftsanbieter unterschiedlich viele Merkmale auf. Für die 2438 hessischen Beherbergungsbetriebe im „HRS Holidays – das Ferienhausportal“ wurden zusammen insgesamt 1398 unterschiedliche Unterkunftseigenschaften gefunden. Die nachfolgende Abbildung 3 zeigt die 18 am häufigsten angegebenen Unterkunftseigenschaften.

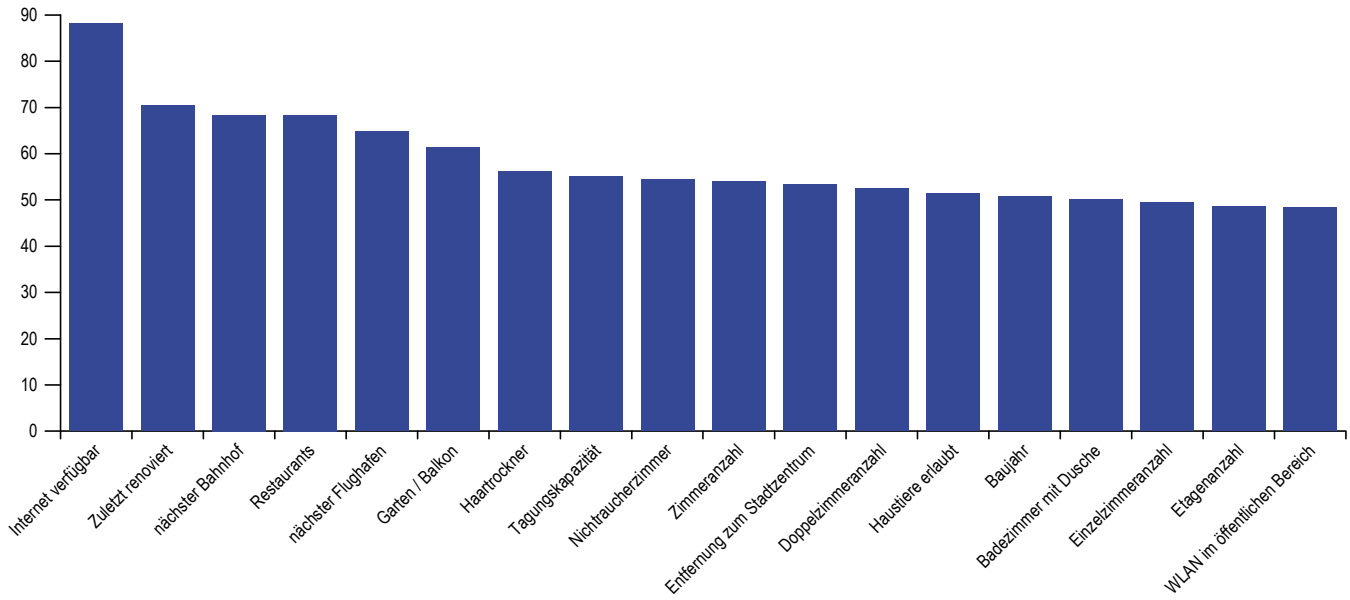
Bei fast allen Beherbergungsbetrieben (88,2 %) wurde verfügbares Internet angegeben. 70,5 % wiesen den Zeitpunkt der letzten Renovierung aus. 68,4 % gaben die Entfernung zum nächsten Bahnhof, 64,8 % zum nächsten Flughafen und 53,4 % zum Stadtzentrum an. 54,0 % stellten die Gesamtanzahl der Zimmer, 52,5 % die Anzahl der Doppelzimmer und 49,5 % die Anzahl an Einzelzimmern dar. Zu den am häufigsten genannten Unterkunftseigenschaften gehörten ferner u. a. das Vorhandensein eines Gartens/Balkons (61,3 %), die Verfügbarkeit von Haartrocknern (56,2 %), das Vorhandensein und die Kapazität von Tagungsräumen (55,1 %), die Mitnahmemöglichkeit für Haustiere (51,5 %), das Baujahr (50,7 %) oder die Etagenanzahl (48,5 %).

Zimmer- und Bettenanzahl

Ein zentrales Merkmal der Tourismusstatistik ist, neben der angebotenen Zimmeranzahl der Beherbergungsbetriebe, die Anzahl der verfügbaren Betten. Diese wird bislang auskunftspflichtig erhoben. Die Gesamtzahl der angebotenen Betten (je Betrieb) war im „HRS Holidays – das Ferienhausportal“ sowie in anderen nach verfügbaren Merkmalen manuell inspizierten Online-Portalen (z. B. Expedia.de, Trivago.de, Booking.com) nicht abrufbar. Bei etwa 1318 (54,1 %) Betrieben waren jedoch sowohl die Anzahl der insgesamt angebotenen Zimmer, als auch jeweils die der Einzel- oder Doppelzimmer verfügbar. Die übrigen 1120 (45,9 %) Betriebe wiesen keine Angaben bezüglich der Zimmeranzahl auf.

Angenommen wurde, dass die Angabe der Zimmeranzahl von der Betriebsgröße abhängt. Da außer der Zimmer- oder Etagenanzahl jedoch keine Eigenschaften verfügbar waren, die auf die Betriebsgröße hingewiesen haben, wie etwa Fläche in m², Stärke des Hotelpersonals, der Umsatz, das Vorhandensein eines eigenen Parkhauses, die Anzahl der Restaurants, das Vorhandensein eines

Abbildung 3: Am häufigsten verwendete Unterkunftseigenschaften hessischer Beherbergungsbetriebe auf dem HRS-Holiday-Portal (in %)

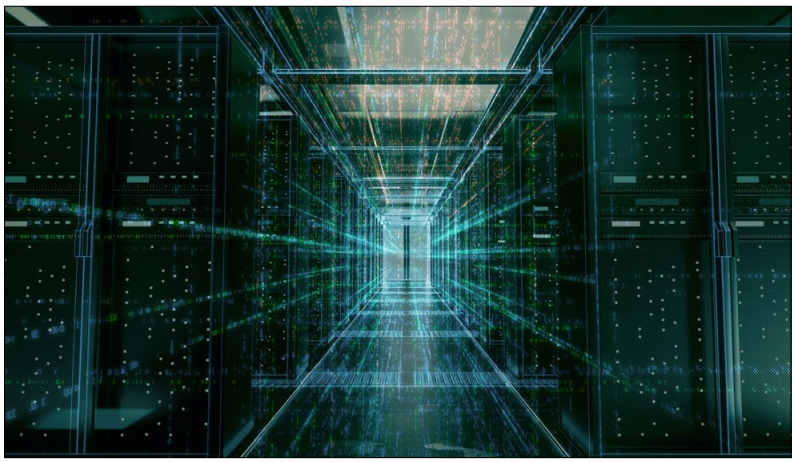


eigenen Wellnessbereichs oder Zugehörigkeit zu einer Kette, konnte dieses Merkmal nicht zur Erklärung herangezogen werden, warum im einen Fall die Angabe der Zimmeranzahl vorlag und im anderen Fall nicht. Als weiteres Erklärungsmerkmal für die Angabe der Zimmeranzahl wurde die den Betrieb charakterisierende Unterbringungsart untersucht. Diese war als Filtermöglichkeit auf der Hauptwebseite verfügbar. Es handelte sich jedoch nicht um eineindeutige Angaben, da die Beherbergungsbetriebe gleichzeitig mehrere Unterbringungsarten (bspw. Hotel, All-Inclusive-Hotel, Ferienhaus, Ferienwohnung, Familiär geführtes Hotel, Feriendorf oder Bauernhof) als für den eigenen Betrieb gültig erklären konnten. Ein Zusammenhang zwischen einer bestimmten Unterbringungsart und der fehlenden Zimmeranzahl konnte nicht gefunden werden. Betriebe mit Zimmeranzahl haben die eigene Unterbringung im Mittel jedoch mit 9,4 Merkmalen beschrieben, während die Betriebe ohne Zimmeranzahl durchschnittlich nur 3,9 Merkmale aufwiesen.

Für die 1318 Betriebe mit ausgewiesener Zimmeranzahl wurde mithilfe der folgenden Annahmen hinsichtlich der Bettenanzahl pro Zimmertyp, die Anzahl verfügbarer Betten pro Betrieb ermittelt:

- Annahme 1: Ein Einzelzimmer enthält eine Schlafgelegenheit.
- Annahme 2: Ein Doppelzimmer enthält zwei Schlafgelegenheiten.

Die Schwierigkeit bestand nun darin, dass sich die beiden Zimmermerkmale Einzelzimmer und Doppelzimmer bezogen auf die jeweiligen Ausprägungen nicht strikt separat betrachten ließen. Ein Doppelzimmer konnte auch als Einzelzimmer genutzt werden. Entsprechend konnte dasselbe Zimmer gleichzeitig sowohl als Einzel- oder Doppelzimmer auf der Webseite ausgewiesen werden. Deshalb waren neben Betrieben, bei denen sich die verschiedenen Zimmerarten korrekt zur angegebenen Gesamtzahl an Zimmern aufaddieren ließen, auch Unterkünfte feststellbar, bei denen dies nicht möglich war. Die Summe der Anzahl aus Einzel- und Doppelzimmern konnte bspw. größer sein als die Anzahl aller Zimmer, da die Betreibenden einen Teil der Zimmer oder alle Zimmer gleichzeitig als Einzel- und Doppelzimmer anbieten konnten. Angenommen ein Betrieb mit



© Connect world – Fotolia.com

einer angegebenen Gesamtzahl von insgesamt fünf Zimmern verfügt über fünf Doppelzimmer, die auf dem Buchungsportal ebenfalls als fünf Einzelzimmer angeboten werden, so wäre die ermittelte Bettenanzahl (fünf Einzelzimmer je eine Schlafgelegenheit + fünf Doppelzimmer je zwei Schlafgelegenheiten = 15 Schlafgelegenheiten) um mindestens fünf Betten überschätzt worden. In dem Fall, dass die Beherbergungsbetriebe neben den angegebenen Einzel- und Doppelzimmern noch über andere Zimmerkategorien (Suiten oder Apartments) in ihrem Bestand verfügten, konnte die Summe aus Einzel- und Doppelzimmern kleiner sein als die Gesamtzahl an Zimmern. Bei einer solchen Konstellation drohte die Unterschätzung der angebotenen Schlafgelegenheiten.

Bei 70,1 % der Betriebe mit ausgewiesenen Zimmeranzahlangaben stimmte die Gesamtanzahl der Zimmer mit der Summe der Anzahl aus Einzel- und Doppelzimmern überein. In diesem Fall ergab sich die geschätzte Anzahl angebotener Schlafgelegenheiten durch das Addieren von jeweils zwei Schlafgelegenheiten pro Doppelzimmer und einer Schlafgelegenheit pro Einzelzimmer gemäß den Annahmen.

Bei 7,1 % der Betriebe mit ausgewiesenen Zimmeranzahlangaben war die Gesamtanzahl der Zimmer kleiner als die Summe der Anzahl aus Einzel- und Doppelzimmern. In diesem Fall wurde die Bettenanzahl zunächst nur über die Doppelzimmeranzahl mit jeweils zwei Schlafgelegenheiten und die dann auftretende Differenz zur Gesamtzimmeranzahl mit jeweils einer Schlafgelegenheit bestimmt und addiert. Es wurde somit angenommen, dass die Zimmer mit unbekannter Katego-

Tabelle 1: Lage- und Streuungsparameter von Zimmern und Betten hessischer Beherbergungsbetriebe im HRS-Holiday-Portal

Art der Angabe	Zimmer	Doppelzimmer	Einzelzimmer	Betten
min	1	1	1	2
max	1 008	587	587	1 661
unteres Quartil	18	10	5	28
Median	30	17	9	49
oberes Quartil	61	36	19	98
Mittelwert	31,94	18,64	10,12	56,31
Standardabweichung	21,53	13,15	8,18	43,34

rie (Differenz der Doppelzimmer zur Gesamtanzahl von Zimmern) jeweils über mindestens eine Schlafgelegenheit verfügten.

Bei 19,3 % der Betriebe mit ausgewiesenen Zimmeranzahlangaben war die Gesamtanzahl der Zimmer größer als die Summe der Anzahl aus Einzel- und Doppelzimmern. In diesem Fall wurde die Bettenanzahl zunächst über die Doppelzimmeranzahl mit jeweils zwei Schlafgelegenheiten und die Einzelzimmeranzahl mit jeweils einer Schlafgelegenheit bestimmt. Die dann noch auftretende Differenz zur Gesamtzimmeranzahl wurde mit jeweils einer Schlafgelegenheit berücksichtigt und addiert.

Die beschriebenen Fälle machten 97,0 % der Beherbergungsbetriebe mit Angaben zur Zimmeranzahl aus. Es verblieben noch 3,0 % mit Angaben zur Zimmeranzahl, jedoch ohne Angaben zur Art der Zimmer. Für diese Fälle wurden die benötigten Werte mit einem Imputationsverfahren geschätzt. Dabei wurden in einem ersten Schritt ausreißerbereinigte Mittelwerte der Anzahl an Doppel- und Einzelzimmern (siehe Tabelle 1) zur Berechnung entsprechender Anteile an der Summe aus den beiden Zimmerarten verwendet.

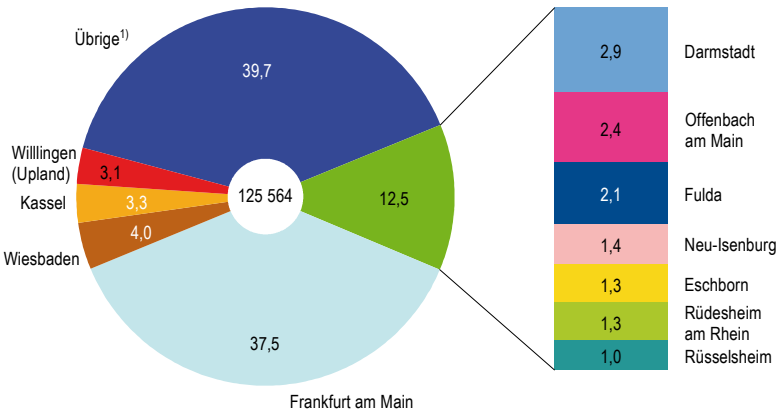
Dabei hatten hessische Beherbergungsbetriebe im Mittel etwa 19 Doppelzimmer und 10 Einzelzimmer. Die geschätzten Anteile der beiden Zimmerarten betrugen für Einzelzimmer 35,2 % und für Doppelzimmer 64,8 %.

In einem zweiten Schritt wurde die verfügbare Zimmeranzahl der Betriebe ohne Angaben zur Zimmerart über die berechneten Anteile auf die nicht beobachtbaren Zimmerarten aufgeteilt. Die Anzahl der Betten dieser 39 Betriebe wurde anschließend nach den Annahmen (zwei Schlafgelegenheiten pro Doppelzimmer und eine Schlafgelegenheit pro Einzelzimmer) berechnet.

Die Ergebnisse der geschätzten Bettenanzahlen hessischer Beherbergungsbetriebe im „HRS Holidays - das Ferienhausportal“ verteilt auf hessische Städte und Gemeinden sind in Abbildung 4 enthalten.

Insgesamt wurden im „HRS Holidays - das Ferienhausportal“ 125 564 Betten angeboten. Frankfurt am Main war mit einem Anteil von 37,5 % mit deutlichem Abstand die Stadt in Hessen mit dem größten Bettenangebot. Der Unterschied zu der in Abbildung 2 dargestellten regionalen Verteilung der hessischen Beherbergungsbetriebe war die deutlichere Konzentration des Bettenangebotes. Frankfurt am Main stellte mehr als ein Drittel aller in Hessen über das „HRS Holidays – das Ferienhausportal“ angebotenen Betten. Somit waren die meisten und die größten Beherbergungsbetriebe in Frankfurt am Main ansässig. Die Städte und Gemeinden mit jeweils weniger als 1,0 % des Bettenangebots stellten zusammen 39,7 % aller angebotenen Betten. Diese Kategorie war um 24,4 Prozentpunkte kleiner als bei der Betrachtung hessischer Beherbergungsbetriebe, bei der die Bettenanzahl nicht berücksichtigt wurde. Die zweitgrößte hessische Stadt in Bezug auf das ermittelte Bettenangebot war mit 4,0 % Wiesbaden. Bei der reinen Betrachtung der Verteilung der Beherbergungsbetriebe im „HRS Holidays – das Ferienhausportal“ lag Wiesbaden auf Platz 8 von 10. Eine Schlussfolgerung könnte sein, dass Wiesbaden eher größere Beherbergungsbetriebe beheimatete als bspw. Kassel mit 3,3 % oder

Abbildung 4: Bettenangebot hessischer Beherbergungsbetriebe im HRS-Holiday-Portal nach Städten und Gemeinden (in %)



1) Städte und Gemeinden, die jeweils weniger als 1,0 % der hessischen Schlafgelegenheiten aufwiesen.

Willingen (Upland) mit 3,1 %. Die weiteren hessischen Städte und Gemeinden mit dem größten Bettenangebot im HRS-Holiday-Portal waren Darmstadt (2,9 %), Offenbach am Main (2,4 %), Fulda (2,1 %), Neu-Isenburg (1,4 %), Eschborn sowie Rüdesheim am Rhein (jeweils 1,3 %) und Rüsselsheim (1,0 %). Im Vergleich zu den im statistischen Bericht der hessischen Tourismusstatistik von September 2018 ausgewiesenen 264 808 Schlafgelegenheiten wurde mit dem Scraping des HRS-Portals mit 125 564 Betten etwa die Hälfte der über den herkömmlichen Beschaffungsweg erhobenen Bettenanzahl erreicht.

Die Verteilung des hessischen Bettenangebotes im „HRS Holidays – das Ferienhausportal“ nach Hotelkategorie (HRS-Sterneanzahl) in Abbildung 5 zeigt, dass die verfügbaren Betten zum größten Teil durch 4-Sterne-Unterkünfte (45,2 %) gestellt wurden.

Darauf folgten die 3-Sterne-Unterkünfte (38,6 %), die 5-Sterne-Unterkünfte (7,0 %), die 2-Sterne-Unterkünfte (6,3 %), die Unterkünfte ohne Stern (2,1 %) und die 1-Sterne-Unterkünfte (0,8 %). Die mittlere und die gehobenen Hotelkategorien dominierten damit mit 83,8 % das ermittelte hessische Bettenangebot.

Umgebungsinformationen

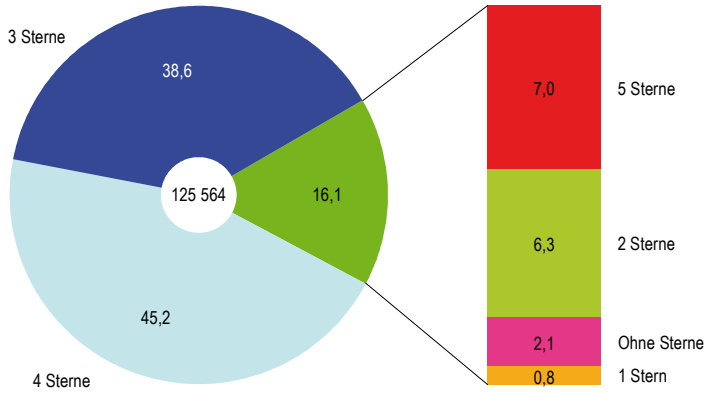
Für viele Reisende können die Ausstattung eines Hotels, die Verpflegung, die Zimmergröße, das Freizeitangebot vor Ort aber insbesondere für Dienst- und Geschäftsreisende der Zugang zur lokalen Infrastruktur sehr wichtig sein.

Im „HRS Holidays – das Ferienhausportal“ war das Merkmal „Entfernung zum nächsten Bahnhof“ bei 68,4 % der Unterkünfte verfügbar. 53,4 % gaben die Entfernung zum Stadtzentrum an. Für 64,8 % der hessischen Unterkünfte waren Angaben zum nächsten Flughafen und für 48,4 % zur nächsten Autobahn verfügbar. Angaben zum nächsten Krankenhaus waren bei 13,1 % der Beherbergungsbetriebe verfügbar.

Eine Darstellung der mittleren Entfernungen zu den genannten Infrastrukturobjekten findet sich in Abbildung 6.

Im Mittel war das Stadtzentrum 1,68 km von den hessischen Beherbergungsbetrieben entfernt.

Abbildung 5: Geschätztes Bettenangebot hessischer Beherbergungsbetriebe im HRS-Holiday-Portal nach Hotelkategorien (Sterneanzahl) (in %)

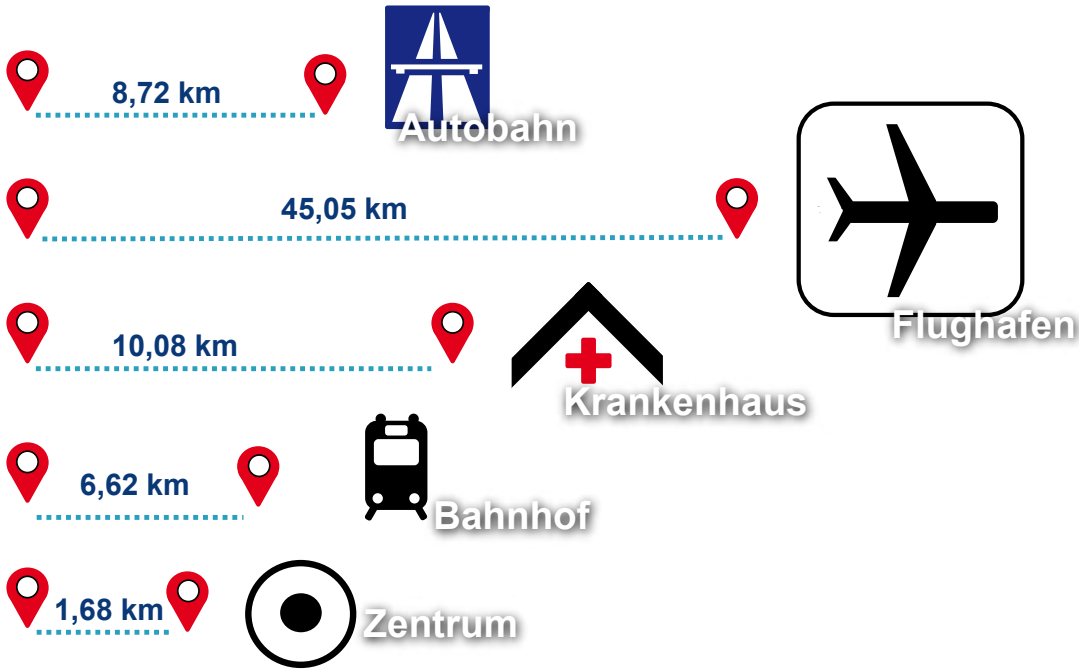


Der nächste Bahnhof war nach 6,62 km und die nächste Autobahn nach 8,72 km zu erreichen. Das nächste Krankenhaus war 10,08 km, der nächste Flughafen 45,05 km entfernt. Dieses Beispiel zeigt, dass auch diese Umgebungsinformationen statistisch auswertbar sind.

Fazit

Das Scrapen von hessischen Beherbergungsbetrieben und die Extraktion von Unterkunftsmerkmalen im HRS-Holiday-Portal verlief reibungslos. Nach manueller Kontrolle stellte sich heraus, dass alle auf der Webseite enthaltenen Unterwebseiten der Beherbergungsbetriebe gefunden und abgespeichert worden sind. Ein sorgfältiges Anpassen der „regulären Ausdrücke“ und das Nutzen der CSS- und Xpath-Selektoren führten dazu, dass die Unterkunftsinhalte korrekt zugeordnet und ausgewiesen worden sind. Die Vielfalt und Menge der gewonnenen Daten und insbesondere die große Anzahl an Beherbergungsbetrieben verdeutlichen das große Potenzial des Scrapings von Online-Portalen, um digitale Daten zu gewinnen. Als nächster Schritt ist geplant, weitere kommerzielle Online-Portale der Beherbergung wie Trivago oder Booking.com automatisiert auszulesen und auszuwerten. Eine Herausforderung wird zum einen darin bestehen, die Informationen aus den jeweiligen Teilbetrachtungen mit den Informationen im statistischen Unternehmensregister und der Beherbergungsstatistik zusammenzuführen. Zum anderen wird es herausfordernd, Überschneidungen bzw. Mehrfachzählungen herauszufiltern sowie die Qualität der Angaben und des berechneten Angebots an Schlafgelegenheiten

Abbildung 6: Mittlere Entfernung hessischer Beherbergungsbetriebe zu ausgewählten Infrastrukturobjekten in HRS-Holiday-Portal



beurteilen zu können. Der Abgleich mit den Daten der amtlichen Beherbergungsstatistik und dem statistischen Unternehmensregister wird zudem zeigen, ob das Verfahren dazu geeignet ist, den Erhebungsaufwand beim Statistikproduktionsprozess zu verringern.

Ausblick: Webscraping in der amtlichen Statistik, national und international

Webscraping ist außerhalb der amtlichen Statistik eine etablierte Technik zur Suche und Strukturierung von öffentlich zugänglichen Informationen aus dem Internet. Die amtliche Statistik nutzt diese Art der Informationsgewinnung bislang nur vereinzelt, um eigenständig Daten zu gewinnen (z. B. Preise im Internet), Auskunftspflichtige zu entlasten oder die eigentliche Erhebung zu unterstützen. Daneben wird der mögliche Einsatz für neue Zwecke weiter untersucht. Der europäische Verbund nationaler statistischer Ämter ESS hat dem Webscraping in einem von der Europäischen Kommission geförderten Forschungsprojekt („ESSnet Big Data“, 2016–2018) zwei eigene Arbeitspakete gewidmet (Webscraping von Online-Stellenbörsen, Webscraping von Unternehmenseigenschaften). Im anschließenden Forschungsprojekt „ESSnet Big Data II“ (2018–2020) sollen diese beiden Arbeitspakete weitergeführt werden, um konkrete Verfahren zur Statistikproduktion zu implementieren. Zusätzlich dazu nimmt im Arbeitspaket „Innovative Datenquellen für die Tourismusstatistik“ u. a. das Webscraping eine zentrale Stelle ein. Dabei wird auf das Webscraping von Buchungs- und Bewertungsportalen besonders Wert gelegt. Darüber hinaus sind weitere Datenquellen wie Wetterdaten, anonymisierte Mobilfunkdaten, Verkehrsdaten und viele mehr in der Diskussion. Das Hessische Statistische Landesamt ist als einer der deutschen Partner an diesem Arbeitspaket beteiligt, und zwar mit den Schwerpunkten „Webscraping“ und „Nutzung anonymer Mobilfunkdaten“.

Normen Peters; Tel.: 0611 3802-517
E-Mail: normen.peters@statistik.hessen.de

Literaturverzeichnis

Hessisches Statistisches Landesamt, 2018. Gäste und Übernachtungen im hessischen Tourismus im September 2018. Statistischer Bericht, Kennziffer: G IV 1 - m 09/2018 [online], [Zugriff am: 20.09.2018]. Verfügbar unter: https://statistik.hessen.de/sites/statistik.hessen.de/files/GIV1m_18-09.pdf

PETERS, Normen, 2018. Webscraping von Unternehmenswebseiten und maschinelles Lernen zum Gewinnen von neuen digitalen Daten. Hessisches Statistisches Landesamt. Sonderveröffentlichung [online], [Zugriff am: 20.09.2018]. Verfügbar unter: https://statistik.hessen.de/sites/statistik.hessen.de/files/Webscraping_von_Unternehmenswebseiten.pdf

Statistisches Bundesamt, 2018. Monatserhebung im Tourismus. Qualitätsbericht, Fachserie 6 Reihe 7.1 - August 2018 [online], [Zugriff am: 20.09.2018]. Verfügbar unter: https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/BinnenhandelGastgewerbeTourismus/Tourismus.pdf;jsessionid=76A45A9E5B0D7ABD06521840E30381D8.InternetLive2?__blob=publicationFile