universität wien

MAGISTERARBEIT

Titel der Magisterarbeit

Nowcasting mit Hilfe von Internet-Suchstatistiken

Verfasst von

Matthias Schmidl, Bakk.rer.soc.oec.

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften
(Mag.rer.soc.oec.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 066 913
Studienrichtung lt. Studienblatt:	Magisterstudium Volkswirtschaftslehre
Betreuerin / Betreuer	Univ.-Prof. Dipl.-Ing. Dr. Robert Kunst

Abstract

Diese Arbeit untersucht das Potenzial von Internet-Suchdaten um kurzfristige Prognosen unterschiedlicher österreichischer Wirtschaftsdaten zu verbessern. Dazu werden Google Trends-Daten zu monatlichen Google-Indizes (GI) aggregiert und mittels schrittweiser Modellselektion ein favorisiertes Modell aus einem Set möglicher Modell-Kandidaten bestimmt. In den Nowcasting-Experimenten werden Out-of-Sample Forecasts von autoregressiven Benchmark-Modellen mit Nowcasts von erweiterten GI-Modellen verglichen. Die Anwendungsdaten stammen aus den Bereichen Arbeitsmarkt und Tourismus, wobei speziell die Arbeitslosen- und Jugendarbeitslosenquote sowie die Ankünfte von deutschen und niederländischen Gästen in Österreich untersucht werden. Die Ergebnisse zeigen, dass mit GI erweiterte Modelle gegenüber den Benchmark-Modellen in allen vier Anwendungsfällen einen geringeren RMSE und MAE in den Out-of-Sample Forecasts produzieren.

Abstract

This work examines the potential of internet search query data to improve short-term forecasts of various Austrian economic data. Therefore, Google Trends-data are aggregated to monthly Google-Indices (GI) and a stepwise model selection algorithm determines a prefered model out of a set of candidate models. In nowcasting experiments out-of-sample forecasts of autoregressive benchmark models are compared to nowcasts of models augmented with the GI. The data is taken from the tourism sector and the labor market. In particular unemployment and youth unemployment data, as well as arrivals of German and Dutch visitors in Austria are analysed. The results show that GI augmented models produce a lower out-of-sample RMSE and MAE in all four applications compared to the benchmark models.

Keywords: Google Trends, forecasting, nowcasting, unemployment, tourism

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

1 Einleitung

Das Internet spielt in den letzten Dekaden eine zunehmend wichtige Rolle im Alltag der Menschen. Insbesondere stellt es häufig einen Ausgangspunkt ökonomischer Aktivitäten dar: Von der Informationssuche nach Produkten und deren Einkauf, der Planung und Buchung von Urlauben bis zur Jobsuche und Bewerbung kann mittlerweile alles online erledigt werden. Eine große Rolle dabei haben Suchmaschinen, welche das World Wide Web durchforsten und ihre NutzerInnen zu den gewünschten Seiten und Informationen führen. Dabei offenbaren die NutzerInnen ihre wahren Absichten – sei es durch die Suche nach Konzertkarten, Zugverbindungen oder Bekleidungsgeschäften. Das Suchverhalten kann demnach Hinweise über zukünftige Aktivitäten der NutzerInnen geben.

Das Wissen über die (geplanten) Aktivitäten der InternetnutzerInnen eröffnet sowohl in der wissenschaftlichen Forschung als auch in der Wirtschaft eine Vielzahl an Anwendungsmöglichkeiten. Im Fokus der gegenständlichen Arbeit steht dabei weniger die Vorhersage zukünftiger Entwicklungen (*predicting the future*), sondern vielmehr die Bestimmung des gegenwärtigen Zustands (*predicting the present*). Von Interesse ist diese Möglichkeit aufgrund der Tatsache, dass amtliche Statistiken in der Regel mit zeitlicher Verzögerung veröffentlicht werden, Internet-Suchstatistiken hingegen beinahe in Echtzeit verfügbar sind.

In der Folge kann die gegenwärtige Entwicklung von makroökonomischen Daten noch vor der Veröffentlichung der Daten auf Basis der eigenen Vergangenheit und unter zusätzlicher Zuhilfenahme von Internet-Suchstatistiken geschätzt werden. Zu beachten ist dabei, dass die Internet-Suchstatistiken bereits für den zu prognostizierenden Zeitraum verfügbar sind. Im Zusammenhang mit Forecasts dieser Art, welche auch Daten aus dem zu prognostizierenden Zeitraum berücksichtigen, wird häufig der Begriff „*Nowcasting*“ verwendet.

Mit dem Start des Produkts *Insights for Search*[1] bietet *Google* die Möglichkeit, Statistiken über das Suchaufkommen der *Google*-NutzerInnen abzurufen. Erste Nowcasting-Anwendungen damit zeigen Choi & Varian (2009a,b), welche die Genauigkeit von ökonometrischen Modellen in mehreren Fällen verbessern können. Aufbauend auf diesen und weiteren Studien steht die Anwendung von *Google*-Suchdaten in Modellen zur kurzfristigen Prognose österreichischer Wirtschaftsdaten im Mittelpunkt dieser Arbeit.

---

[1] Seit 2012 in *Google Trends* integriert.

Im Speziellen werden dabei die Arbeitslosen- und Jugendarbeitslosenquote sowie die Ankünfte von deutschen und niederländischen Gästen in Österreich prognostiziert.

Ziel der gegenständlichen Arbeit ist es zu untersuchen, ob mit *Google*-Suchdaten erweiterte Modelle gegenüber Benchmark-Modellen eine höhere Prognosegenauigkeit erzielen. Dazu werden *Google Trends*-Daten zu monatlichen Google-Indizes (GI) aggregiert und einem Baseline-Modell hinzugefügt. Ein schrittweiser Modellselektionsalgorithmus bestimmt dann ein favorisiertes Modell aus einem Set möglicher Modell-Kandidaten. Um die Prognosegenauigkeit der Modelle zu evaluieren, werden die Out-of-Sample Forecasts der Baseline-Modelle mit Nowcasts von erweiterten GI-Modellen verglichen.

Zunächst wird jedoch die Bedeutung des Internets am Arbeitsmarkt und für den Tourismus beleuchtet und jeweils eine Literaturübersicht zu Anwendungen von *Google*-Suchdaten in diesen Bereichen gegeben. Anschließend wird die Datengrundlage bestehend aus den *Google Trends*-Daten, Arbeitsmarktdaten und der Beherbergungsstatistik beschrieben. Danach wird die gewählte Methodik bezüglich der Punkte Datenanalyse, Modellierung und Evaluation im Detail erläutert. Die empirische Analyse der Anwendungsbeispiele erfolgt schließlich im darauffolgenden Abschnitt.

2 Hintergrund

In diesem Abschnitt werden jeweils Hintergrundinformationen zur Entwicklung und Bedeutung des Internets für den Tourismus und am Arbeitsmarkt gegeben. Anschließend werden die Ergebnisse aktueller wissenschaftlicher Arbeiten mit *Google Trends*-Daten in der Arbeitsmarkt- sowie in der Tourismusstatistik erörtert. Zunächst wird jedoch der Begriff „*Nowcasting"* und dessen Verwendung in der Ökonomie erläutert.

2.1 Der Begriff "Nowcasting" in der Ökonomie

Nowcasting ist ein Mischwort aus den Begriffen *Now* und *Forecasting* und beschreibt in der Ökonomie die Prognose der Gegenwart beziehungsweise der unmittelbaren Zukunft oder Vergangenheit.[2] Die Relevanz von Nowcasts ergibt sich aufgrund der Tatsache, dass grundlegende volkswirtschaftliche Kennzahlen zumeist mit zeitlicher Verzögerung ausgewiesen werden und eventuell einer späteren Revision unterzogen werden. Zur frühzeitigen Bestimmung des gegenwärtigen Zustands einer ökonomischen Variable werden auf Basis der bis dahin verfügbaren Informationen Schätzungen vorgenommen. Charakteristisch für diese, als Nowcasts bezeichnete, Schätzungen ist die Verwendung von zum Zeitpunkt *t* bereits ausgewiesenen Daten zur Bestimmung einer ebenfalls in *t* indizierten, jedoch noch nicht ausgewiesenen Zielvariable. Im Unterschied dazu verwenden Forecasts die zum Zeitpunkt *t* verfügbaren Informationen, um Prognosen über zukünftige Werte der Zielvariable (für die Zeitpunkte $t+i$, $i=1,2,\ldots$) zu erstellen. Nowcasts werden in einer Vielzahl an Institutionen, insbesondere in Wirtschaftsforschungsinstituten und Zentralbanken, zum Monitoring des gegenwärtigen Zustands der Volkswirtschaft eingesetzt und gewinnen dafür zunehmend an Bedeutung.

2.2 Bedeutung des Internets am Arbeitsmarkt

Die steigende Popularität des Internets und die stetig wachsende Zahl der BenutzerInnen ist mitverantwortlich für tiefgreifende Veränderungen am Arbeitsmarkt – insbesondere im Recruitment beziehungsweise bei der Jobsuche. Laut einer empirischen Untersuchung von Weitzel et al. (2011) zu Recruiting Trends der „Top-1.000 Unternehmen Deutschlands" sowie „der Top-300-Unternehmen aus den Branchen

[2] vgl. Bańbura et al. (2013)

Finanzdienstleistung, IT und Öffentlicher Dienst“ werden 87% der vakanten Stellen auf unternehmenseigenen Webseiten und 61% auf Internet-Jobbörsen ausgeschrieben. Schließlich resultieren rund 72% aller tatsächlichen Einstellungen aus den genannten Kanälen. Generell sind Onlinekanäle das meistgenutzte Mittel zur aktiven Stellensuche, wie eine Untersuchung einer Online-Umfrage mit 10.050 TeilnehmerInnen zeigt [Weitzel et al. (2014)]. Laut dieser Umfrage liegen Internet-Stellenbörsen (von rund 65% der Befragten genutzt) in der Liste der beliebtesten Kanäle vor Unternehmens-Webseiten (37%) und Karrierenetzwerken (34%).

Diese Tendenzen am Arbeitsmarkt spiegeln sich in der Folge ebenso im Datenverkehr des Internets wider. Ein Abbild davon verzeichnen die diversen Suchmaschinen, einerseits direkt über die aktive Jobsuche der NutzerInnen und andererseits indirekt über die Informationssuche nach karrierererelevanten Themen. So könnte beispielsweise auch das Suchaufkommen von Begriffen betreffend der Arbeitslosenversicherung einen Hinweis auf die Entwicklung der Arbeitslosenrate geben. Die Anwendung von Suchmaschinenstatistiken zur Echtzeitbestimmung der Arbeitsmarktentwicklung ist folglich ein logischer Schritt in der modernen Wirtschaftsforschung. Um den derzeitigen Stand der wirtschaftswissenschaftlichen Forschung darzulegen, werden im nächsten Abschnitt aktuelle Arbeiten zu dieser Thematik erläutert.

2.3 Literaturübersicht zu Anwendungen in der Arbeitslosenstatistik

Die ersten Arbeiten, die *Google Trends*-Daten[3] mit Arbeitsmarktdaten verknüpften, wurden ein Jahr nach der Veröffentlichung von *Google Insights for Search* im Sommer 2009 publiziert. Nach eigener Recherche können aktuell elf wissenschaftliche Arbeiten auf diesem Gebiet gezählt werden.

Pionierarbeit leisten dabei Askitas & Zimmermann (2009) in ihrer explorativen Studie zur deutschen Arbeitslosenquote, die eine starke Korrelation zwischen *Google Trends*-Daten und der Arbeitslosenrate aufzeigt. Dazu bilden die Autoren monatliche *Google*-Indikatoren für vier Gruppen von Suchbegriffen, indem sie jeweils die ersten beiden Monatswochen sowie die dritte und vierte Monatswoche aggregieren. Mit Hilfe dieser Konstruktion versuchen die Autoren zu klären, ob sich die *Google Trends*-Daten der

[3] Die Literaturübersicht beschränkt sich auf Publikationen, in welchen Suchmaschinendaten von *Google Trends* angewendet werden, da gegenwärtig keine ähnlichen Arbeiten bekannt sind, die Daten anderer Suchmaschinen berücksichtigen.

ersten Hälfte des laufenden Monats (Woche 1 + 2) oder der zweiten Hälfte (Woche 3 + 4) des Vormonats besser als Prädiktoren für die gegenwärtige Arbeitslosenquote eignen. Schließlich favorisieren sie ein Modell, das *Google*-Daten aus Woche 3 und 4 des jeweiligen Vormonats zu den Suchbegriffen „Arbeitsamt“ oder „Arbeitsagentur“ und diversen deutschen Jobbörsen (unter anderem. Stepstone, Jobworld, Jobscout) beinhaltet.

Forschung im Bereich Suchdaten-basiertes Nowcasting der Arbeitsmarktentwicklung wird insbesondere von Zentralbanken betrieben.[4] Suhoy (2009) untersucht für die Bank of Israel die Eignung von Google-Indizes zum Monitoring der ökonomischen Aktivität, insbesondere während des konjunkturellen Abschwungs 2008. Von sechs ausgewählten Suchkategorien (siehe Abschnitt 3.1.2) erweist sich die Kategorie „Personalwesen (Personalvermittlung und Zeitarbeitsunternehmen)“ als am besten für die Vorhersage der monatlichen Arbeitslosenrate geeignet. Demnach ist eine steigende Häufigkeit von Suchanfragen in dieser Kategorie mit einer Erhöhung der Arbeitslosenrate verbunden.

Chronologisch nach Askitas & Zimmermann (2009) und Suhoy (2009) veröffentlichen Choi & Varian (2009b) ihre Analyse von *Google*-Suchanfragen und Erstanträgen auf Arbeitslosenhilfe (*Initial Claims*) in den USA. Choi und Varian vergleichen darin die Out-of-Sample Forecasts eines Baseline-Modells [AR(1)] mit den Forecasts eines erweiterten Modells, das zusätzlich *Google*-Index-Variablen berücksichtigt. Der monatliche *Google*-Index wird aus den Werten der ersten Monatswochen der Kategorien „Beruf“ und „Sozialwesen und Arbeitslosigkeit“ gebildet. Die Ergebnisse zeigen eine Verbesserung des durchschnittlichen absoluten Fehlers (Mean Absolute Error, MAE) um 12,9% im kurzen und 15,7% im langen Zeitfenster gegenüber dem Baseline-Modell. In einer späteren Version mit aktualisierten Daten [Choi & Varian (2012)] können die Autoren hingegen nur eine geringe Verbesserung des korrigierten Bestimmtheitsmaß (adj. R^2) und keine Verbesserung des Out-of-Sample MAE feststellen. Sie argumentieren jedoch, dass das *Google Trends*-Modell speziell um Wendepunkte verbesserte Vorhersagen liefert und demzufolge bei der Identifikation von Wendepunkten behilflich sein kann.

Eine empirische Untersuchung zur italienischen Arbeitslosenquote legt D’Amuri (2009) vor. Die zu prognostizierende Variable ist dabei die auf Quartalsbasis verfügbare italienische Arbeitslosenquote, welche durch den Italian Labor Force Survey ermittelt

---

[4] Zu den von Zentralbankpersonal durchgeführten Forschungsarbeiten in diesem Bereich zählen die Studien von Suhoy (2009) für die Bank of Israel, D’Amuri (2009) und D’Amuri & Marcucci (2009) für die Banca d’Italia und Chardwick & Sengül (2012) für die Zentralbank der Republik Türkei.

wird. Als erklärende Variable verwendet der Autor *Google Trends*-Daten für den Suchbegriff „Stellenangebote“ („offerte di lavoro“). Diese auf wöchentlicher Basis verfügbaren Internet-Suchdaten aggregiert der Autor durch die Bildung von Vierteljahresdurchschnittswerten zu einem *Google Index*. Im Vergleich zu einem ARIMA(1,1,0)-Benchmark-Modell wird der mittlere quadratische Fehler (MSE) vom erweiterten *Google Index*-Modell halbiert. In einem weiteren Artikel führt D'Amuri zusammen mit Marcucci ähnliche Untersuchungen zur US-amerikanischen Arbeitslosenrate durch [D'Amuri & Marcucci (2009)]. Hier betrachten sie eine Vielzahl an linearen und nicht-linearen Modellen, die zum Teil mit *Google*-Suchstatisiken für den Suchbegriff „jobs“ ergänzt werden oder den Indikator der *Initial Claims* beinhalten. Die Autoren stellen fest, dass die besten Forecasts von Modellen geliefert werden, welche *Google*-Index-Variablen beinhalten. Darüber hinaus übertrifft das beste *Google*-Index-Modell die Forecast-Ergebnisse des *Survey of Professional Forecasters* der Federal Reserve Bank of Philadelphia.

Untersuchungen zu norwegischen Daten führen Anvik & Gjelstad (2010) in ihrer Master-Thesis durch. In Anlehnung an Choi & Varian (2009b) und D'Amuri (2009) vergleichen sie die Vorhersagefehler[5] von *Google*-Index-Modellen mit den eines Baseline-Modells sowie eines Modells mit dem Frühindikator „Veröffentliche Stellenanzeigen“ (*published job advertisements*). Insgesamt verwenden sie vier *Google*-Indizes, die *Google Trends*-Daten für korrespondierende Suchbegriffe zu bestimmten Gruppen zusammenfassen: „Arbeitslosigkeit und Sozialleistungen“, „Arbeitsämter und Institutionen“, „private Jobvermittler“ und „aktive Suche“. Die *Google Trends*-Daten aggregieren sie zu Monatsdaten, indem sie Durchschnittswerte für die Zeiträume von 15. bis zum 14. des Folgemonats berechnen. Ihre Ergebnisse zeigen eine Steigerung der Vorhersagegenauigkeit des bestperformenden *Google*-Index-Modells gegenüber dem Baseline-Modell um durchschnittlich 18,3% und gegenüber dem weiteren Frühindikator-Modell um durchschnittlich 16%.

Zu ähnlichen Ergebnissen gelangen Bughin (2011) in seiner Untersuchung zur belgischen und Chadwick & Şengül (2012) zur türkischen Arbeitslosenrate. In beiden Studien können Vorhersageverbesserungen gegenüber Benchmark-Modellen festgestellt werden. Eine Studie von Barreira et al. (2013) gelangt in den Nowcasting-Experimenten zur Arbeitslosenraten und Neuwagenverkaufszahlen in vier Staaten Süd-Westeuropas hingegen zu unterschiedlichen Ergebnissen. Während im Bezug auf die Arbeitslosenrate Portugals, Frankreichs und Italiens Verbesserungen erzielt werden können, gelingt dies für Spanien nicht.

---

[5] Wurzel des mittleren quadratischen Fehlers (RMSE).

Eine empirische Untersuchung zur Jugendarbeitslosigkeit in Frankreich legen Fondeur & Karamé (2013) vor. Die Autoren zeigen darin eine signifikante Verbesserung der Nowcasts der Arbeitslosenrate für die Gruppe der 15-24-Jährigen mit Hilfe von *Google Trends*-Daten, stellen aber auch Vorhersageverbesserungen bei anderen Altersgruppen fest. Die Autoren streichen jedoch insbesondere heraus, dass die Vorhersagegenauigkeit unter Verwendung von *Google*-Suchdaten bei der Jugendarbeitslosenquote am stärksten steigt.

2.4 Bedeutung des Internets im Tourismus

Das Internet spielt auch eine zunehmend wichtige Rolle bei der Planung und Buchung von Reisen und Urlauben. Laut einer von Eurostat 2012 durchgeführten Studie [Seybert (2012)] nutzen rund die Hälfte der InternetuserInnen der EU-27 Webdienste im Zusammenhang mit Reisen (beispielsweise Websites, die Informationen über Unterbringungen, Touristenattraktionen oder Flugpläne bieten oder Buchungsservices für Reisetickets und Hotelzimmer). Ein ähnliches und detaillierteres Bild zeigt die umfangreiche Untersuchung zur Internetnutzung von Reisenden von Fesenmaier et al. (2009) im Auftrag der U.S. Travel Association. Demnach nutzen 2009 über 57% der amerikanischen InternetuserInnen das Internet bei der Reiseplanung und mehr als drei Viertel davon planen Vergnügungsreisen online. Dabei zählen Online-Reisebüros, Suchmaschinen, Unternehmenswebseiten und Webseiten des Reiseziels zu den meistverwendeten Webservices. Eine spezifische Analyse des Kommunikations- und Informationsprozesses von Österreich-UrlauberInnen nimmt die Österreich Werbung (2013) in der Studie „Customer Journey Online – Österreich Urlauber“ vor. Deutsche und heimische Österreich-UrlauberInnen verwenden das Internet in der Reisevorbereitung häufig als Informations- und Inspirationsquelle. Im Bezug darauf erreichen Webseiten von Unterkünften, Orten und Regionen die höchsten Vertrauenswerte unter den Befragten. Das persönliche Gespräch in Reisebüros suchen Österreich-UrlauberInnen in der Vorbereitung hingegen weniger häufig als UrlauberInnen generell (4% beziehungsweise 15% der Befragten). Ferner bucht auch nur ca. jeder zehnte Österreich-Urlauber/jede zehnte Österreich-Urlauberin ihren Urlaub in Reisebüros. Bei jenen, die ihren Urlaub online buchen, ist die direkte Buchung über die Website der Unterkunft die meistgewählte Methode (71% der Online-Urlaubsbuchungen).

In Summe beweisen die zitierten Studien die steigende Bedeutung des Internets im Informations- und Buchungsprozess von Reisenden. Als Knotenpunkt des World Wide Web führen viele dieser Wege über Suchmaschinen, welche in der Folge ein Abbild der Reisetätigkeiten ihrer UserInnen erfassen (können). Die Anwendung von

Suchmaschinendaten im Monitoring von Fremdenverkehrszahlen steht im Fokus des folgenden Abschnitts.

2.5 Literaturübersicht zu Anwendungen in der Tourismusstatistik

In ihrer wegweisenden Arbeit demonstrieren Choi & Varian (2009a) unter anderem auch Anwendungsbeispiele im Tourismusbereich. Sie verwenden *Google Trends*-Daten mehrerer Staaten, um die monatlichen Besucherankünfte dieser Herkunftsländer in Hong Kong zu schätzen. Neben Wechselkursen und Indikatorvariablen für die olympischen Spiele in Peking 2008 dienen die *Google Trends*-Daten der Subkategorie „Hong Kong" in der Kategorie „Reisedestinationen" als Prädiktoren. Die *Google Trends*-Daten für die Herkunftsländer USA, Kanada, Großbritannien, Deutschland, Frankreich, Australien, Indien und Japan rufen die Autoren über den Einsatz des regionalen Filters von *Google Trends* ab. Eine Varianzanalyse zeigt einen statistisch signifikanten Einfluss[6] der *Google Trends*-Variablen und eine gute Modellanpassung (korrigiertes R^2 von 0,988).

Vorhersagen wichtiger Tourismuskennzahlen für Dubai nehmen Saidi et al. (2010) vor und verwenden dafür unter anderem *Google Trends*-Daten. Die Autoren kommen jedoch zu unterschiedlichen Ergebnissen: Während die Forecasts der Nächtigungszahlen mit Hilfe der *Google*-Daten kein eindeutiges Bild ergeben, erweisen sich die Suchdaten im Bezug auf die Vorhersage von Ankünften am Dubai Airport als äußerst effektiv.

Aufbauend auf Choi & Varian (2009a) entwickeln Gawlik et al. (2011) eine weitere Methode zur Bestimmung der monatlichen Besucherankünfte in Hong Kong. Statt aggregierten Suchdaten der Kategorie „Reisedestinationen" verwenden sie *Google Trends*-Daten zu spezifischen Suchanfragen. Um die relevantesten Suchbegriffe zu bestimmen, verwenden sie einen Variablenselektions-Algorithmus (*Forward Search*), dessen Ergebnisse mittels Kreuzvalidierung evaluiert werden. Die besten Resultate im Bezug auf den relativen Fehler im Testset liefert eine lokal-gewichtetes Regressionsmodell, das sowohl englische als auch chinesische Suchbegriffe berücksichtigt.

In einer weiteren Studie benutzen Pan et al. (2012) *Google*-Suchdaten, um den Bedarf an Hotelzimmern in der US-amerikanischen Stadt Charleston, South Carolina, vorherzusagen. Im Unterschied zu den bisher genannten Studien, die *Google Trends*-Daten zur Bestimmung monatlicher Zielvariablen verwenden, ist die Zielvariable (Hotelzimmerbelegung in Charleston) in dieser Untersuchung ebenso auf wöchentlicher Basis verfügbar. Ihre Resultate zeigen Vorhersageverbesserungen aller *Google*-Modelle

[6] zum 5%–Niveau.

gegenüber den übrigen Modellen. Insbesondere liefert ein simples ARX(1)-Modell die höchste Nowcast-Genauigkeit[7] unter den berücksichtigten neun Modellen.

Der Tourismus ist in vielen Ländern einer der bedeutendsten Wirtschaftsfaktoren. Einige Zentralbanken versprechen sich eine Verbesserung der bisherigen Methoden zum Monitoring der Fremdenverkehrszahlen durch die Berücksichtigung von Suchmaschinendaten. Studien dazu führen unter anderem Artola & Galan (2012) für die spanische und Matsumoto et al. (2013) für die japanische Zentralbank durch. Artola & Galan (2012) analysieren eine spezifische Anwendung von *Google Trends*-Daten für die spanische Tourismuswirtschaft: Die Vorhersage von Ankünften britischer Gäste[8] in Spanien. Die Autoren stellen fest, dass die Verbesserung der Vorhersagegenauigkeit stark von den zugrunde gelegten Benchmark-Modellen abhängt und in diesem Kontext bewertet werden sollte. Trotzdem streichen sie die Anwendungsmöglichkeiten der *Google Trends*-Daten als Frühindikator für britische Touristenströme mit einem zeitlichen Vorsprung von einem Monat hervor. Matsumoto et al. (2013) fokussieren sich besonders auf die Entwicklung der Inanspruchnahmen von Reisedienstleistungen vor und nach dem starken Erdbeben in Japan im März 2011. Laut den Resultaten erweisen sich die *Google*-Suchdaten als nützlich, um die Auswirkungen dieses unerwarteten Schocks quantitativ zu erfassen. Die Autoren sehen die Anwendung von Suchdaten als Komplement zu klassischen ökonomischen Indikatoren, streichen aber auch die Vorteile als Nowcasting-Tool nach Schockereignissen hervor.

---

[7] Im Bezug auf die durchschnittlichen absolute Prozentabweichung und der Wurzel der durchschnittlichen, quadrierten Prozentabweichung.

[8] TouristInnen aus dem Vereinigten Königreich sind Spaniens bedeutendste Touristengruppe.

3 Datengrundlage

In diesem Abschnitt wird die Datenbasis, bestehend aus den *Google Trends*-Daten, den Gästeankünften der TourMIS-Datenbank und der Arbeitslosenstatistik der Statistik Austria beschrieben.

3.1 Google Trends

Die Daten zum Suchverhalten von NutzerInnen der Suchmaschine *Google* (verfügbar unter www.google.com/trends) sind die am häufigsten verwendete Datenquelle für Nowcasting-Aufgaben dieser Art. Um die zuvor zitierten Untersuchungen im Zusammenhang mit der Entwicklung und dem Funktionsumfang von *Google Trends* einordnen zu können, wird in den folgenden Abschnitten der Entwicklungsprozess des Dienstes dargelegt und der aktuelle Funktionsumfang umrissen. Danach wird die Berechnung der *Google Trends*-Daten beschrieben und die Aggregation zu einem monatlichen *Google*-Index erläutert.

3.1.1 Hintergrund

Mit *Google Trends* bietet *Google Inc.* seit 2006 ein Produkt an, mit dessen Hilfe das Suchaufkommen der Suchmaschine *Google* analysiert werden kann. Der Dienst ist seit dem Start einem laufenden Entwicklungsprozess unterzogen und wurde seither in Aufbau und Funktionalität mehrmals modifiziert und erweitert. Darum soll zunächst ein Überblick über die Entwicklungshistorie und Zusammenhänge mit anderen *Google*-Diensten verschafft werden, um anschließend den aktuellen Funktionsumfang zu erläutern, sowie sich daraus ergebende Anwendungsmöglichkeiten aufzuzeigen.

Am 10.Mai 2006 wurde *Google Trends* im Rahmen des jährlichen *Google „Press Day“* der Öffentlichkeit präsentiert.[9] Es ermöglicht die graphische Ausgabe und den Vergleich des Suchvolumens von beliebigen Begriffen der *Google*-Suchmaschine. Bereits zu diesem Zeitpunkt war sowohl eine geographische als auch eine zeitliche Differenzierung der Ergebnisse möglich. Der Entwicklung von *Google Trends* vorrausgegangen ist das Produkt *Google Zeitgeist*, welches eine Auswertung der am häufigsten gesuchten Begriffe eines Jahres erlaubt. *Google Zeitgeist* wird seit der Präsentation des Tools im Dezember 2001 jährlich für das jeweils vergangene Jahr veröffentlicht. Mit *Google Hot*

[9] Mayer, M. (2006).

Trends wird am 22. Mai 2007 eine konzeptionell ähnliche Erweiterung von *Google Trends* vorgestellt, die es erlaubt, die derzeit aktivsten Suchanfragen anzuzeigen Zeitgleich stellt *Google* die bis dahin auch wöchentlich kompilierten *Zeitgeist*-Listen ein.[10]

Am 6. August 2008 startet *Google* mit *Insights for Search* eine Erweiterung von *Google Trends*, die im Funktionsumfang eine deutlich detailliertere Analyse des Suchaufkommens ermöglicht[11]. Einerseits kann damit im Vergleich zu *Google Trends* eine feinere geographische Differenzierung vorgenommen werden und mit kategorischen Filtern gibt es die Möglichkeit, Suchbegriffe nach ihrem kontextuellen Zusammenhang zu analysieren (siehe Abschnitt 3.1.2). Weiters werden in *Insights for Search* Prognosen zur weiteren, kurzfristigen Entwicklung der Suchanfragen angezeigt und eine Downloadfunktion ermöglicht den Export der Daten.

2008 startet *Google* das Projekt *Google Flu Trends*[12] mit dem Ziel, ein Echtzeit-Monitoring des Auftretens von Grippeerkrankungen zu ermöglichen. Ginsberg et al. (2009) zeigen, dass die Ausbreitung von Grippeerkrankungen in den USA mit Hilfe von *Google Trends*-Daten mit hoher Genauigkeit geschätzt werden kann. *Google Flu Trends* ist mittlerweile für 29 Staaten verfügbar und ein ähnliches Projekt zur Denguefieberausbreitung[13] gibt es für 10 Staaten. Ausgehend von diesen Projekten, die reale Daten mit *Google Trends*-Daten verbinden, entwickelt *Google* das Produkt *Google Correlate*, welches am 25. Mai 2011 vorgestellt wurde[14]. Mit *Google Correlate* kann zu beliebigen wöchentlichen oder monatlichen Zeitreihen eine Liste der am besten korrelierenden *Google Trends*-Zeitreihen ausgegeben werden.[15] Schließlich vereint *Google* am 27. September 2012 die beiden Dienste *Trends* und *Insights for Search* unter einer neuen gemeinsamen Oberfläche, die fortan unter *www.google.com/trends* abrufbar ist.

---

[10] Vickrey, C. (2007).

[11] Helft, M. (2008).

[12] http://www.google.org/flutrends/intl/de/.

[13] http://www.google.org/denguetrends/intl/de/.

[14] Mohebbi, M. (2011).

[15] Mohebbi et al. (2011).

3.1.2 Aktueller Funktionsumfang von Google Trends

Der folgende Abschnitt soll eine Einführung in den Aufbau und die Funktionen von *Google Trends* geben. Insbesondere wird beschrieben, wie *Google Trends*-Zeitreihen ausgewählt und für die weitere Verwendung heruntergeladen werden können. Alle für diese Arbeit relevanten Funktionen sind unter https://www.google.com/trends/explore[16] abrufbar.

Grundsätzlich gibt es im „Erkunden"-Bereich (*explore*) drei Möglichkeiten, das Suchaufkommen verschiedener Begriffe zu analysieren: Der Vergleich mehrerer Suchbegriffe miteinander, der Vergleich einzelne Suchbegriffe nach unterschiedlichen Standorten oder der Vergleich einzelner Suchbegriffe nach unterschiedlichen Zeiträumen.

Zusätzlich können die zu analysierenden Suchbegriffe auf geographische Regionen, Zeiträume, Kategorien und bestimmte *Google*-Dienste eingeschränkt werden. Der geographische Filter erlaubt die Einschränkung des zu analysierenden Suchaufkommens auf bestimmte Regionen, die in Österreich bis auf Bundesländer-Ebene reichen. Im Zeitfilter können alle beliebigen Zeiträume zwischen 1. Jänner 2004 und dem gegenwärtigen Tag gewählt werden. Je nach der Länge des gewählten Zeitfensters wird die *Google Trends*-Zeitreihe dann in täglicher, wöchentlicher oder monatlicher Form ausgegeben[17]. Der Kategorien-Filter erlaubt die kontextuelle Betrachtung des Suchaufkommens nach bestimmten Kategorien, indem der Datenpool der Suchanfragen auf bestimmte Kategorien reduziert werden kann. So können beispielsweise Suchanfragen nach Fenstern („Windows") von Suchanfragen nach dem Betriebssystem *Windows* unterschieden werden. In der deutschsprachigen Version von *Google Trends* stehen auf der ersten Ebene 25 Kategorien[18] zur Auswahl, in Summe gibt es 1.422 Kategorien und Sub-kategorien[19]. Ein weiterer Filter ermöglicht die Analyse des Suchaufkommens bestimmter *Google*-Produkte (Websuche, Bildersuche, News-Suche, *Google* Shopping oder *Youtube*-Suche)

---

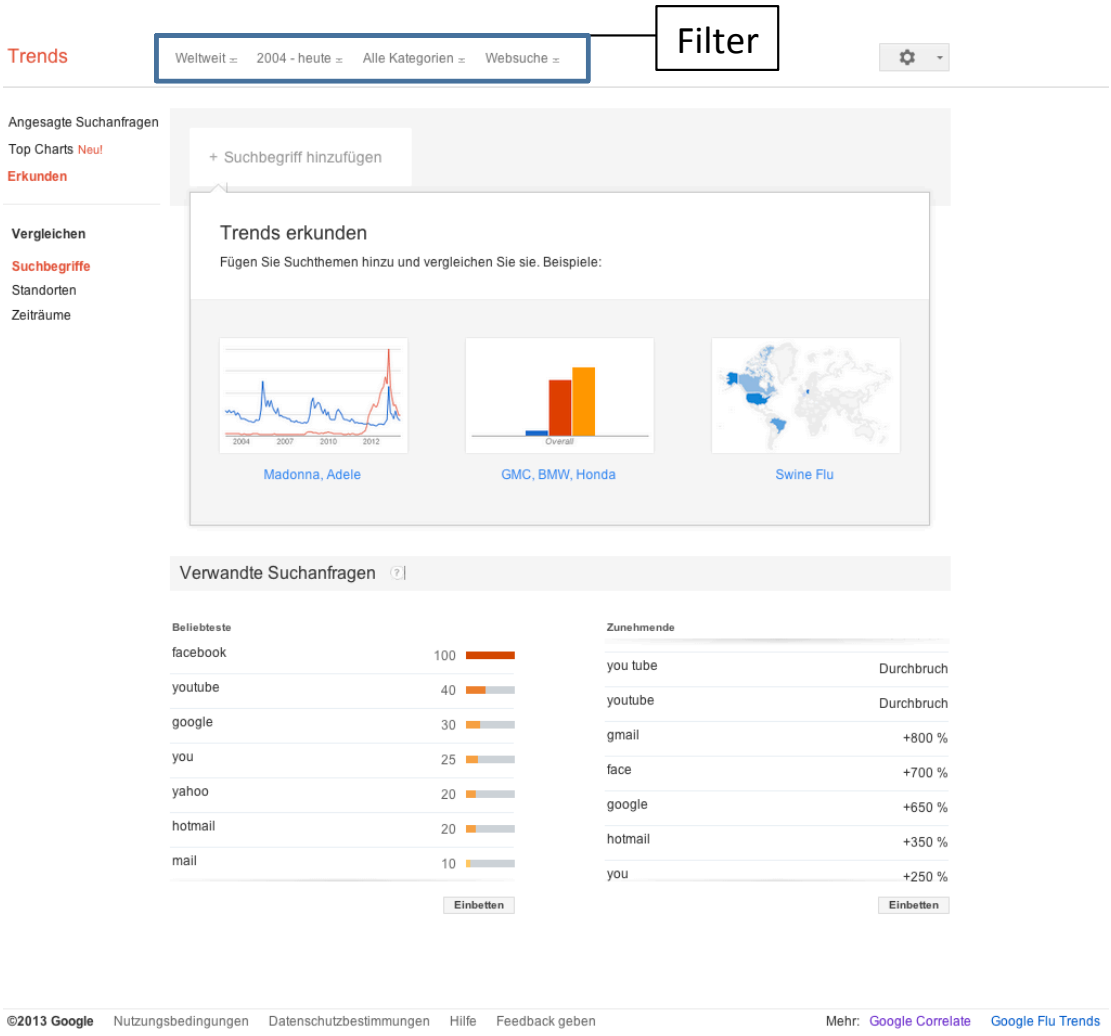
[16] Neben dem Bereich „Erkunden" sind unter den Punkten „Angesagte Suchanfragen" und „Top Charts" auch Funktionen und Technologien der Dienste *Google Trends*, *Google Hot Trends* und *Google Zeitgeist* verfügbar.

[17] Eine genaue Abgrenzung, ab welchem Zeitfenster tägliche, wöchentliche oder monatliche Daten ausgegeben werden, wird in der *Google Trends*-Hilfe nicht erläutert. Jedoch liefern kürzere Zeitfenster (weniger als 90 Tage) eher tägliche Daten und längere Zeitreihen können in den meisten Fällen als wöchentliche Daten heruntergeladen werden.

[18] Eine Liste dieser Kategorien befindet sich unter Appendix 1.

[19] Doppelzählungen sind möglich.

Abbildung 1: Aufbau und Funktionen von Google Trends



Quelle: http://www.google.com/trends/explore

Mit der Eingabe von Suchbegriffen in die entsprechenden Felder werden mehrere Graphiken und Listen angezeigt. Im Feld „Interesse im zeitlichen Verlauf“ wird das skalierte normierte Suchaufkommen[20] als Zeitreihe angezeigt. Zusätzlich werden an verschiedenen Punkten der Zeitreihe korrelierende Ereignisse angemerkt. Bei Auswahl einer Kategorie kann die Entwicklung des Suchaufkommens zu diesem Begriff im Verhältnis zum Suchaufkommen der gesamten Kategorie angezeigt werden.

In einem weiteren Bereich der Seite wird das „regionale Interesse“ nach dem Suchbegriff anhand einer Heatmap und einer korrespondierenden Liste der geographischen Regionen angezeigt. In der Heatmap werden die Regionen nach Suchhäufigkeit farblich markiert, wobei helle Farben ein geringes Suchaufkommen und dunklere Farben ein

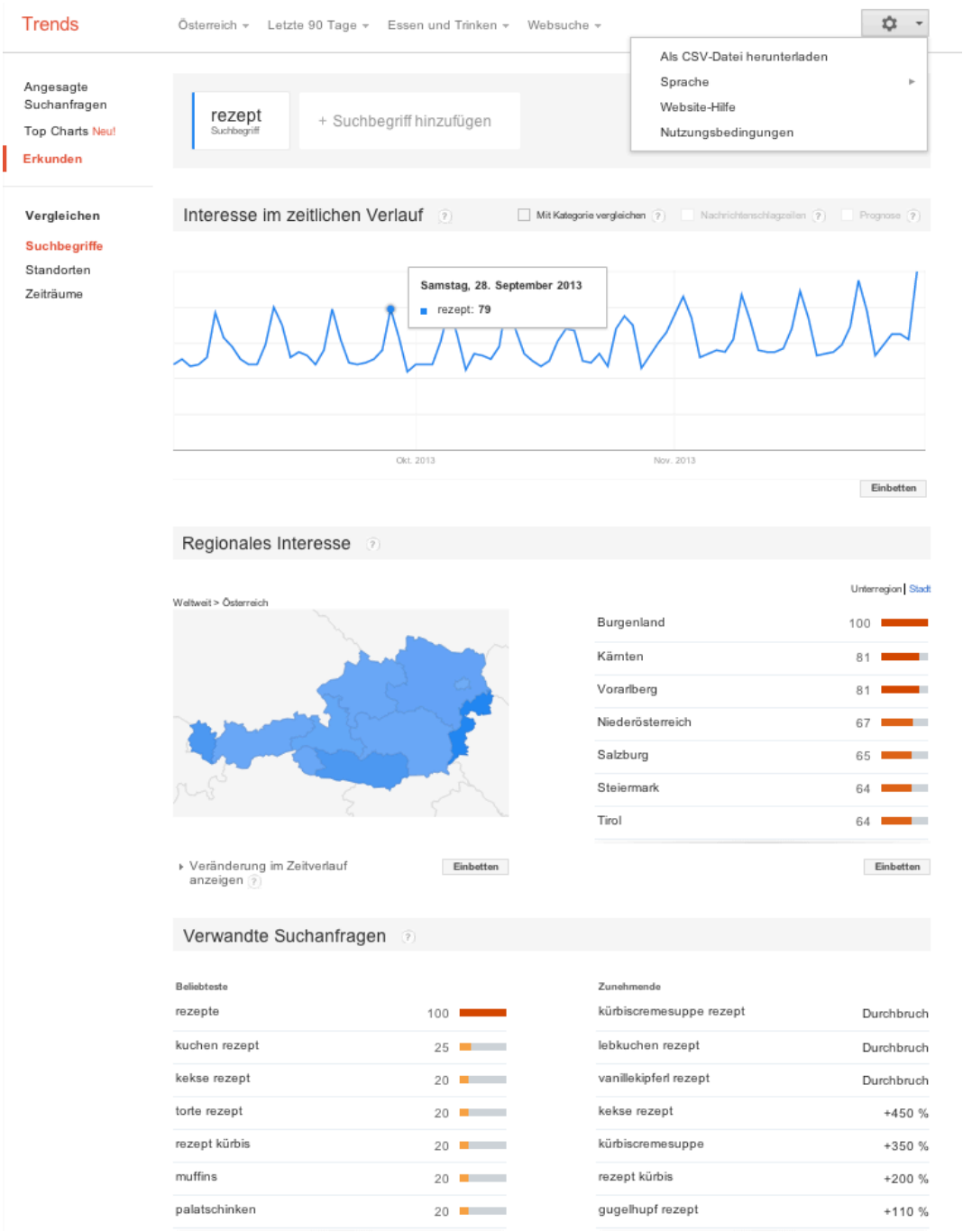
[20] siehe Abschnitt 3.1.3.

hohes Suchaufkommen in der Region signalisieren. In einer nebenstehenden Liste werden die entsprechenden Regionen und Werte nach der Höhe des Suchaufkommens in absteigender Reihenfolge angeführt. Optional kann die Veränderung des regionalen Interesses im Zeitverlauf angezeigt werden. Die Heatmap kann dazu für verschiedene Zeitpunkte im gewählten Zeitraums angezeigt und in animierter Form abgespielt werden.

Im untersten Bereich von *Google Trends* zeigen zwei Listen die beliebtesten sowie am stärksten zunehmenden „verwandten Suchanfragen" an. Als „verwandte Suchanfragen" werden alle Suchanfragen betrachtet, die unter die gewählten Filter fallen beziehungsweise mit dem eingegebenen Suchbegriff korrelieren. Wird kein Suchbegriff eingegeben, zeigt *Google Trends* unter den „beliebtesten Suchanfragen" die häufigsten Suchanfragen in der gewählten Kombination aus Region, Zeitraum, Kategorie und *Google*-Dienst an. Unter den „zunehmenden Suchanfragen" werden die am stärksten im Vergleich zum vorangegangenen Zeitraum[21] wachsenden Suchanfragen angezeigt, welche den gegebenen Einschränkungen genügen. Wird in *Google Trends* ein Suchbegriff eingegeben, reduziert sich der Datenpool, aus dem die Listen erstellt werden, zusätzlich auf ähnliche und korrelierende Begriffe. Bei einer Steigerung der Suchhäufigkeit des gewählten Begriffs um mehr als 5.000 % zeigt *Google Trends* anstelle des tatsächlichen Prozentwerts das Wort „Durchbruch" (*Breakout*) an.

[21] Vormonat, Vorjahr oder Daten aus dem Jahr 2004. Für weiter Information siehe https://support.google.com/trends/answer/94793?hl=de&ref_topic=19360.

Abbildung 2: Eingabebeispiel in Google Trends



Quelle: Google Trends (http://www.google.com/trends)

Unter Voraussetzung eines vorhandenen Google-Accounts können die ausgegebenen Daten zur weiteren Verwendung als CSV-Datei heruntergeladen werden. Die CSV-Datei beinhaltet die Daten zum Suchinteresse im zeitlichen Verlauf (gegebenenfalls auch mit der Entwicklung im Vergleich zur Kategorie), die Top-Unterregionen und Top-Städte sowie die Listen der häufigsten „verwandten Suchanfragen“ und der zunehmenden „verwandten Suchanfragen“.

3.1.3 Ermittlung der Google Trends-Daten

Die Berechnung der *Google Trends*-Daten erfolgt auf Basis einer Zufallsstichprobe aller Suchanfragen in der *Google-Websuche* beziehungsweise für den entsprechenden *Google*-Dienst. Grundsätzlich geben *Google Trends*-Daten einen normalisierten Anteilswert der Häufigkeit von eingegebenen Begriffen in Relation zur Gesamtzahl der *Google*-Suchanfragen an.

Das Suchvolumen zu beliebigen Suchanfragen wird in Form eines Index ausgewiesen, der „die Wahrscheinlichkeit eines beliebigen NutzerInnens an einem bestimmten Standort in einem festgelegten Zeitraum nach einem gewissen Begriff zu suchen“ [22] angeben soll.

Die Bestimmung des *Google Trends*-Index wird folgenderweise beschrieben:[23] Die Berechnung erfolgen auf Basis einer Zufallsstichprobe der gesamten *Google*-Suchanfragen. Die Analyse kann auf bestimmte Regionen, Zeiträume und Suchkategorien beschränkt werden (siehe Abschnitt 3.1.2). Die Datensätze werden durch das gesamten *Google*-Suchaufkommen in der betreffenden Region normalisiert *(1)*, wodurch der Index trotz unterschiedlich hoher Zahl an Anfragen für einen bestimmten Begriff auch für unterschiedliche Regionen vergleichbar ist. Danach werden die Daten auf einer Skala von 0 bis 100 skaliert *(2)*. Dies geschieht in dem jeder Datenpunkt durch den höchsten normalisierten Wert beziehungsweise durch 100 dividiert wird.

$$Normalisierter\ Wert = \frac{Suchvolumen\ des\ gewählten\ Begriffs}{Gesamtes\ Suchvolumen} \quad (1)$$

$$Skalierter\ Wert = \frac{Normaliserter\ Wert}{Höchster\ normalisierter\ Wert} \cdot 100 \quad (2)$$

Die daraus resultierenden Suchstatistiken reichen bis zum 1. Jänner 2004 zurück und sind standardmäßig als Zeitreihe auf wöchentlicher Basis und für kurze Zeiträume auf täglicher Basis verfügbar. In einzelnen Fällen mit niedrigem Suchvolumen wird der *Google* Index auch als monatliche Zeitreihe angegeben.

[22] Support.google.com (2007).

[23] zum Folgenden: ebd.

3.1.4 Transformation der Google Trends-Daten

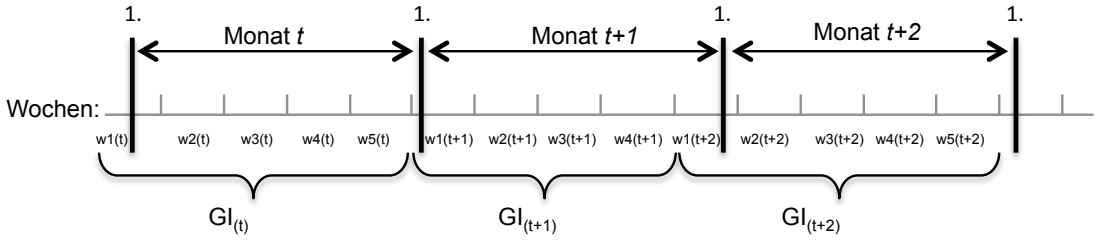
In diesem Abschnitt wird die Transformation der wöchentlichen *Google*-Daten in einen monatlichen *Google Index* (abgekürzt GI) erläutert. Diese Notwendigkeit ergibt sich aufgrund der Tatsache, dass die nachfolgend verwendeten, realen Anwendungsdaten aus verschiedenen Quellen stammen und nur in monatlicher Form verfügbar sind.

Um die wöchentlichen *Google Trends*-Daten als monatliche Zeitreihe verwenden zu können beziehungsweise sie zu einer monatlichen Zeitreihe zu aggregieren, gibt es mehrere Möglichkeiten: Choi & Varian (2009a,b) verwenden in ihren Berechnungen ausschließlich die erste Monatswoche. Askitas und Zimmerman (2009) bilden jeweils aus den ersten beiden Monatswochen sowie der dritten und vierten Monatswoche Durchschnittswerte. In dieser Arbeit wird ein ähnlicher Ansatz zu D'Amuri (2009) verfolgt:

Für die Berechnung des GI werden jeweils die Wochen des Monatsersten als Split-Kriterium herangezogen. Der GI für das Monat *t* ist dann der Durchschnittswert der Wochen des Monatsersten bis zur Vorwoche des darauffolgenden Monatsersten. Beispielsweise liegt der Monatserste im Juni 2013 in der 22. Kalenderwoche (Sonntag), im Juli in der 27. Kalenderwoche (Montag). Der GI für Juli 2013 wird nun als das arithmetische Mittel aus den Werten der 22., 23., 24., 25. und 26. Kalenderwoche gebildet. Für das gegenwärtige Monat wird der GI als Durchschnitt der bis dahin verfügbaren Wochenwerte berechnet. Es kann also bereits ein GI für dieses Monat berechnet werden, obwohl das Monat noch nicht vorüber ist.

Mit dieser Methode können die Daten des gegenwärtigen Monats unter Berücksichtigung von Daten aus dem gleichen Zeitraum um bis zu acht Wochen vor den üblichen Veröffentlichungsterminen der offiziellen Daten (häufig am Monatsende des darauffolgenden Monats) geschätzt werden. Die geringfügige Verschiebung der „GI-Monate" gegenüber realer Monate kann darüber hinaus dadurch begründet werden, dass zwischen der Informationssuche via Internet und den tatsächlich gesetzten Handlungen oft ein zeitlicher Abstand besteht. Insbesondere werden gegen Monatsende durchgeführte Suchanfragen häufig erst im darauffolgenden Monat in realen Handlungen umgesetzt.

Abbildung 3: Zeitstruktur des Google-Index



Quelle: Eigene Abbildung

Im Unterschied zu D'Amuri (2009) wird nicht die Woche des Monatszwölften als Split-Kriterium herangezogen, sondern der intuitivere Zugang über die Woche des Monatsersten gewählt. Weiters verwendet D'Amuri (2009) jeweils die vorangegangenen vier Wochen zum Monatszwölften zur Berechnung des GI, wodurch verfügbare *Google Trends*-Daten in einigen Monaten unberücksichtigt bleiben. Mit der hier gewählten Methodik werden hingegen sämtliche Daten verwendet.

3.2 Arbeitslosenstatistik

Die amtliche Arbeitslosenstatistik der *Statistik Austria* beruht auf der Mikrozensus-Arbeitskräfteerhebung und wird regelmäßig rund 30 Tage nach Ende des Referenzmonats veröffentlicht. Die Ergebnisse werden sodann an *Eurostat* übermittelt und in dessen statistische Plattform gespeist[24].

Für die Mikrozensuserhebung[25] wird quartalsmäßig eine geschichtete Zufallsstichprobe (nach NUTS-2-Regionen) von 23.000 Haushalten befragt, wobei ausgewählte Haushalte über einen Zeitraum von fünf Quartalen beobachtet werden. Die Befragung wird nach einem Rotationsprinzip vorgenommen, sodass in jedem Quartal ein Fünftel der Haushalte zum ersten Mal, ein weiteres Fünftel zum zweiten Mal, usw. befragt wird. Die harmonisierte monatliche Arbeitslosenquote wird dann gemäß ILO[26]-Definition als

$$Arbeitslosenquote = \frac{U_{alter \in [15,74]}}{U_{alter \in [15,74]} + E_{alter \in [15,74]}} \quad (9)$$

[24] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes.

Die in dieser Arbeit verwendeten Arbeitsmarktstatistiken stammen ausschließlich aus dieser Quelle.

[25] zum Folgenden: Gumprecht et al. (2011).

[26] International Labour Organization.

berechnet, wobei $U_{alter \in [15,74]}$ die Zahl der Arbeitslosen und $E_{alter \in [15,74]}$ die Zahl der Erwerbstätigen für die Altersgruppe der 15- bis 74-Jährigen beschreibt. Zur Berechnung der Jugendarbeitslosenrate werden hingegen nur die Zahlen der 15 bis 24-Jährigen berücksichtigt. In der gegenständlichen Untersuchung werden die saisonal nicht bereinigten Originalwerte verwendet, da auch *Google* keine saisonale Anpassung der *Google Trends*-Daten vornimmt.

3.3 TourMIS-Daten

Das Touristische Marketinginformationssystem (abgekürzt TourMIS) liefert die Datenbasis für die spätere Analyse der Entwicklung der Ankunftszahlen in Österreich. TourMIS ist ein von der Abteilung für Tourismus und Hotel Management der MODUL University Vienna, dem Institut für Tourismus und Freizeitwirtschaft der Universität Wien entwickeltes Internetportal, das gemeinsam von Partnern aus der Tourismusbranche, Ministerien und Ämtern entwickelt und gewartet wird.[27]

Grundlage für die via TourMIS verfügbare Ankunftsstatistik ist die Beherbergungsstatistik der Statistik Austria. Diese basiert auf Daten von ca. 75.000 gewerblichen und privaten Beherbergungsbetrieben aus rund 1.600 Berichtsgemeinden. Die Veröffentlichung der monatlichen Ankunfts- und Nächtigungsstatistik erfolgt jeweils ab dem 20. des Folgemonats, spätere Revisionen sind jedoch möglich. Die Ergebnisse können nach Herkunftsland, Destination, Unterkunftsart und Berichtsmonat ausgewertet werden. In späterer Folge werden die Ankunftszahlen für alle bezahlten Unterkunftsarten (Hotels, Ferienwohnungen und Ferienhäuser, Campingplätze, Kinder- und Jugenderholungsheime, Gästehäuser und Herbergen, Kur- und Erholungsheime, Sanatorien, Heil- und Pflegeanstalten) aus den Herkunftsländern Deutschland und Niederlande verwendet.

[27] Eine Auflistung der beitragenden Unternehmen und Institutionen findet sich unter http://www.tourmis.info/cgi-bin/tmintro.pl?action=2&sprache=TXD (abgerufen am 04. September 2013).

4 Methodik

In diesem Abschnitt werden die ökonometrische Vorgangsweise und Modellierungsstrategien diskutiert, sowie die angewendeten Evaluierungsmethoden erläutert.

4.1 Datenanalyse

Die Anwendung eines autoregressiven Modells bedingt eine zugrundeliegende, *stationäre* Zeitreihe. Die Zeitreihe der Zielvariable wird darum anhand ihrers Graphen sowie der Autokorrelationsfunktionen (*ACF, PACF*) analysiert und ggf. transformiert (logarithmiert , differenziert). Die Nullhypothese eines stochastischen Prozesses mit Einheitswurzel wird mittels Augmented Dickey-Fuller-Test (ADF-Test) überprüft[28]. Die Anzahl der darin berücksichtigten AR-Komponenten *k* wird durch $k = \left\lfloor (n-1)^{\frac{1}{3}} \right\rfloor$ bestimmt, wobei *n* die Anzahl der Daten bezeichnet.[29] Für die in dieser Arbeit verwendeten Datenreihen ergibt sich für das Trainingsset (2004 bis 2012; 96 Beobachtungen) somit eine Ordnung von *k*=4.

4.2 Modellierung

In Anlehnung an Choi & Varian (2009a,b) wird ein *autoregressives* Baseline-Modell[30] (*AR Modell*) mit Modellen, die zusätzlich *Google Index*-Variablen berücksichtigen, verglichen. Im Baseline-Modell M_0 wird die Zielvariable des gegenwärtigen Monats (y_t) durch autoregressive Terme ($y_{t-1},\ldots,y_{t-p}$) bestimmt – der Zielwert wird also ausschließlich auf Basis der vergangenen Entwicklung prognostiziert („*Forecasting*“). Für die Auswahl des Baseline-Modells werden AR-Modelle bis zur Ordnung 12 herangezogen, wobei jenes mit dem niedrigsten AICc-Wert gewählt wird. Darüber hinaus werden die Residuen auf Autokorrelation und Normalverteilung überprüft.

---

[28] Getestet wird mittels der Funktion *adf.test* aus dem R-Package *tseries.*

[29] Die Bestimmung des Default-Werts mit $\left\lfloor (n-1)^{\frac{1}{3}} \right\rfloor$ entspricht der vorgeschlagenen oberen Grenze, mit der sich die Anzahl der AR-Komponenten im generellen ARMA(p,q)-Setup korrespondierend zur Stichprobengröße *n* erhöht. [Trapletti & Hornik (2013) nach Banerjee et al. (1993) und Said & Dickey (1984)].

[30] Alternierend als Baseline-Modell, Vergleichsmodell und Benchmark-Modell bezeichnet.

Sei M_0 das Baseline-Modell:

$$M_0: \quad y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + e_t \tag{3}$$

Die Variable e_t beschreibt den Fehlerterm. Diesem Modell werden nun bis zur Konvergenz zu einem finalen Modell[31] schrittweise *Google* Index-Variablen ($GI_t^{(1)}, \dots, GI_t^{(m)}$) zum Zeitpunkt *t* hinzugefügt beziehungsweise entfernt.

Sei M_1 das *Google*-Modell:

$$M_1: \quad y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{m} \gamma_j GI_t^{(j)} + e_t \tag{4}$$

Gegenüber dem Baseline-Modell berücksichtigt das *Google*-Modell zusätzlich Daten, die zum prognostizierenden Zeitpunkt *t* bereits ausgewiesen sind. Aufgrund dessen kann hier auch von *Nowcasting* gesprochen werden. Die Methode der automatischen Variablenselektion optimiert die Auswahl der berücksichtigten GI-Variablen in der Hinsicht, sodass ein gewähltes Informationskriterium (AIC, AICc, BIC) größtmöglich reduziert wird. Die theoretischen Hintergründe zur schrittweisen Modellselektion und unterschiedliche Ansätze dazu werden im nachfolgenden Kapitel dargelegt.

4.3 Informationskriterien

Ziel der Modellselektion ist es, ein „bestes“ Modell aus einem Set mehrerer, als Kandidaten in Frage kommender Modelle auszuwählen. Die in der gegenständlichen Arbeit angewandte Methode der schrittweisen Modellselektion nach dem korrigierten Akaike Informationskriterium (*Stepwise Model Selection nach AICc*) verfolgt dabei einen informationstheoretischen Ansatz.[32]

Das von Akaike (1973) vorgeschlagene Informationskriterium („*an information criterion*“) liefert als relatives Qualitätsmaß von statistischen Modellen eine Grundlage zur Modellselektion. Das auch als *Akaike information criterion, AIC* bekannte Informationskriterium ist definiert als

$$AIC = -2 \log L + 2K \tag{5}$$

wobei *L* die Likelihood und *K* die Anzahl der Parameter des Modells ist. Das AIC gibt einen Tradeoff zwischen Goodness-of-Fit und Über-/Unterparametrisierung wider: der

[31] Alternierend als erweitertes Modell, *Google*-Modell oder GI-Modell bezeichnet.

[32] zum Folgenden: Burnham & Anderson (2002).

erste Teil der rechten Seite der Formel (*-2*log L*) tendiert mit wachsender Parameterzahl kleiner zu werden, wohingegen letzterer Term mit der Anzahl der Parameter wächst ("*Strafterm*"). Da der Strafterm unabhängig von der Stichprobengröße ist neigt das AIC bei großer Parameterzahl relativ zur Stichprobengröße eher Modelle mit verhältnismäßig vielen Parametern zu liefern. Um den sich daraus ergebenden Bias zu korrigieren entwickelte Sigura (1978) eine Variante des AIC (abgekürzt: AICc) für kleinere Stichprobengrößen. Das AICc ist definiert als

$$AICc = AIC + \frac{2K(K+1)}{n-K-1} = -2\log L + 2K + \frac{2K(K+1)}{n-K-1} \quad (6)$$

Das AICc besitzt gegenüber dem AIC einen zusätzlichen Bias-Korrektur-Term für kleine Stichprobengrößen *n* relativ zur Parameterzahl *K*. Ist das Verhältnis *n/K* jedoch groß, wird der Einfluss des Korrekturterms geringer, sodass AIC und AICc dazu tendieren, dieselben Modelle zu liefern. Generell empfehlen Burnham & Anderson (2002) aus diesem Grund die Verwendung des AICc in Fällen eines niedrigen Stichprobengrößen-Parameter-Verhältnisses[33].

Das von Schwarz (1978) vorgeschlagene Bayessche Informationskriterium (BIC oder Schwarz-Bayes Criterion, SBC) verwendet ebenfalls einen von der Stichprobengröße abhängigen Strafterm. Das BIC ergibt sich als

$$BIC = -2\log L + \ln(n)\,K \quad (7)$$

und bestraft zusätzliche Parameter gegenüber dem AIC/AICc schärfer. In weiterer Folge liefert das BIC tendenziell niederparametrisierte Modelle als AIC/AICc.

Generell verfolgen AIC/AICc und BIC jedoch unterschiedliche Zielsetzungen: Während mittels AIC/AICc möglichst gute Vorhersagen getroffen werden sollen, setzt das BIC die Existenz des „wahren", datengenerierenden Modells voraus und nähert sich diesem asymptotisch mit Wahrscheinlichkeit 1 an, vorausgesetzt es ist im Datensatz enthalten. Auf Grundlage dieser Fakten wird im Folgenden das AICc als Informationskriterium der Wahl herangezogen.

[33] Als Schwelle wird *n/K* < 40 angegeben.

4.4 Stepwise Modelselection nach Venables & Ripley

Die in der gegenständlichen Arbeit verwendete Modellselektionsroutine baut auf der R-Funktion *stepAIC* aus dem Package *MASS* auf. Im Folgenden soll die Funktionsweise des Algorithmus detailliert beschrieben werden.

Als Ausgangspunkt dient ein Model einer geeigneten Klasse (hier lineare Modelle der Klasse *lm*), das auf unterschiedliche Parametrisierungen untersucht wird. Im Argument *scope* kann der zu untersuchende Modellbereich mit einer Liste zweier Objekte des Typs *formula* übergeben werden, wobei die *upper* Komponente das komplexeste und die *lower* Komponente das simpelste Modell definiert. Mit den weiteren Argumenten *k* und *direction* kann der Multiplikator der Freiheitsgrade des Informationskriteriums gewählt beziehungsweise der Modus der schrittweisen Variablenselektion spezifiziert werden.[34] Als Modus kann festgelegt werden, ob der Prozess den Modellen ausschließlich Variablen hinzufügt (*forward*), entfernt (*backward*) oder beides (*both*) durchführt.[35]

Die Funktion *stepAIC* unternimmt nun iterativ folgende Schritte: In jedem Schritt wird der Effekt des Hinzufügens beziehungsweise Entfernens einzelner Modellparameter mit Hilfe der Funktionen *addterm* und *dropterm* untersucht. Beide Funktionen passen alle sich um einen Parameter unterscheidende Modellvarianten an und berechnen jeweils die Veränderung des gewählten Informationskriteriums (welches durch das Argument *k* bestimmt ist)[36]. Nun wird jener Modellparameter dem Ausgangsmodell hinzugefügt beziehungsweise entfernt, der das gewählte Informationskriterium größtmöglich reduziert. Dieses Modell dient im nächstfolgenden Schritt als Ausgangsmodell. Der gesamte Prozess wird im Weiteren so lange wiederholt, bis das Informationskriterium nicht mehr reduziert werden kann.

In der vorliegenden Arbeit wird eine adaptierte Form dieses Algorithmus verwendet. Die Funktionen wurden insofern angepasst, dass die Modellwahl nach dem AICc-Kriterium erfolgt. Die theoretische Grundlage dafür ist die Empfehlung von Burnham & Anderson (2002).

---

[34] Per Default ist mit k=2 das AIC festgelegt, k=log(n) ergibt das BIC.

[35] In dieser Arbeit wird die Variante *both* verwendet.

[36] Die Berechnung des Informationskriteriums erfolgt mittels der Funktion *extractAIC*.

4.5 Evaluierung

Zur Evaluierung der Modelle werden einerseits die zugrundeliegenden Modellannahmen überprüft und andererseits die Nowcasting-Resultate in einer Out-of-Sample Vorhersage miteinander verglichen. Die Verteilungsannahmen der Residuen werden mittels diagnostischer Plots und verschiedener Hypothesentests analysiert. Die Unkorreliertheit der Fehler wird mit dem Breusch-Godfrey-Test der Ordnung 1 und 2 überprüft, die Homoskedastizität mit dem Breusch-Pagan-Test und die Normalverteilung der Residuen mit Shapiro-Wilk-Test.

Um Out-of-Sample Vorhersagen produzieren zu können wird das verfügbare Datenset in ein Training- und ein Testset geteilt, wobei das Trainingsset von 1. Jänner 2004 bis zum 31.Jänner 2011 reicht und in Summe 96 Beobachtungen umfasst. Die Vorhersageergebnisse der Modelle werden sodann mit dem Root-Mean-Square-Error (RMSE) und dem Mean-Absolute-Error (MAE) verglichen. Die Bestimmung des RMSE beziehungsweise des MAE erfolgt mit

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \tag{8}$$

$$MAE = \frac{1}{n} \cdot \sum_{t=1}^{n} |y_t - \hat{y}_t|$$

wobei $\hat{y}_t$ der geschätzten Wert des Modells und y_t der tatsächliche Wert des Testsamples (der Größe *n*) zum Zeitpunkt *t* ist.

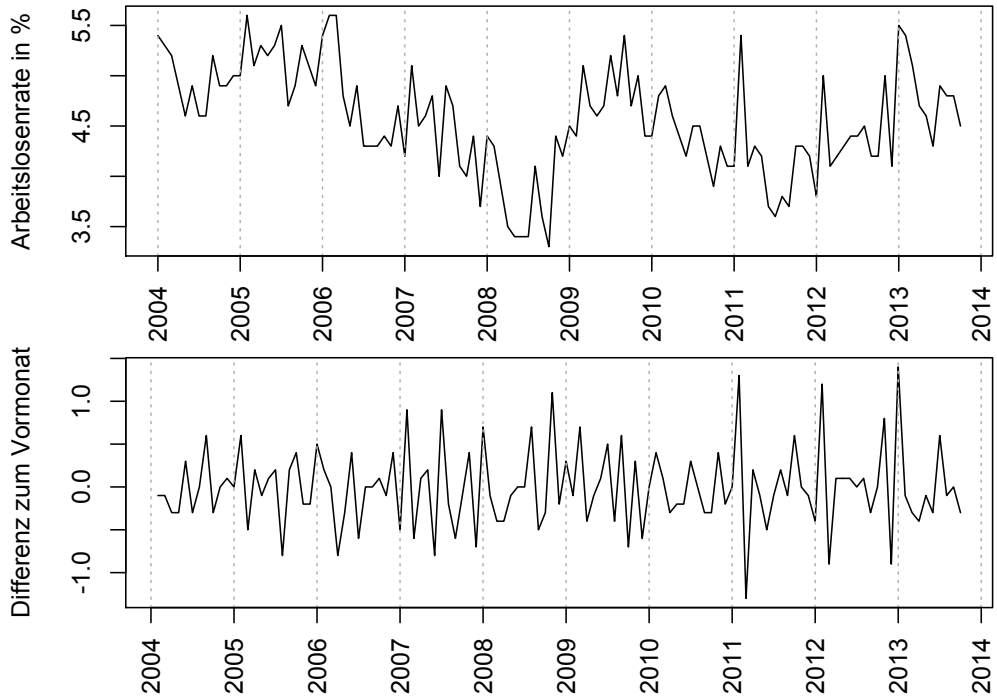
Die Nowcast-Ergebnisse der Modelle werden darüber hinaus graphisch analysiert. Eine direkte Gegenüberstellung der Schätzfehler zeigt, zu welchen Zeitpunkten sich die Ergebnisse unterscheiden und wann Verbesserungen erzielt werden können. Hinweise über die Signifikanz von Vorhersageverbesserungen werden über den Kurvenverlauf der kumulierten Summe der quadrierten Fehler gewonnen. Zeigen die Graphen eine divergierende Entwicklung, kann auf einen signifikanten Unterschied der Vorhersage-ergebnisse geschlossen werden.

5 Empirische Ergebnisse

5.1 Arbeitslosenquote in Österreich

Die Arbeitsloserate in Österreich bewegt sich von 2003 bis 2014 zwischen 3,3 und 5,6%. Um die Stationarität der Zeitreihe zu überprüfen, werden vorerst die Autokorrelationsfunktionen (siehe Appendix 2) untersucht. Die ACF klingt mit zunehmender Lag-Anzahl jedoch nur langsam ab, was auf keine Stationarität der Zeitreihe hindeutet. Auch der ADF-Test (p-Wert = 0,417) verwirft die Nullhypothese einer Einheitswurzel zum 5%-Signifikanzniveau nicht. In der Folge wird mit den Differenzen zum Vormonat weiter verfahren. Diese Vorgehensweise im Zusammenhang mit Arbeitslosenraten wird unter anderem auch in Terasvirta et al. (2010) vorgeschlagen. Die Differenzen scheinen nun eine stationäre Zeitreihe zu bilden – die ACF liegt nach dem ersten Lag konstant innerhalb der Konfidenzbänder[37] und ebenso die PACF zeigt kein nicht-stationäres Muster. Im Fall der differenzierten Zeitreihe liefert der ADF-Test einen Wert von -6,288, welcher signifikant zum 5% Niveau ist.

Abbildung 4: Arbeitslosenquote



Quelle: Eigene Abbildung auf Basis von Eurostat (2013),

[37] entsprechen White Noise.

Auf Grundlage der Differenzen zum Vormonat werden nun lineare kleinst-quadrate Schätzer für unterschiedliche AR-Strukturen berechnet. Modell 2 wird aufgrund des niedrigsten AICc-Wertes (gegenüber den übrigen berücksichtigten Modellen) als Baseline-Modell gewählt. Darüber hinaus erklärt dieses Modell im Vergleich zu den anderen Modellen relativ betrachtet den größten Teil der Varianz, das korrigierte R^2 liegt mit 0,231 am höchsten. Weitere Tests betreffend der üblichen Regressionsanahmen (siehe Appendix 3) legen keine Ablehnung der getroffenen Modellwahl nahe. Sowohl der Breusch-Godfrey-Test auf Autokorrelation der Residuen der Ordnung 1 und 2, der Breusch-Pagan-Test auf Homoskedastie als auch der Shapiro-Wilk-Test auf Normalverteilung der Residuen legen keine Ablehnung der zugrundeliegenden Modellannahmen zum 5%-Signifikanzniveau nahe.

Dem Baseline-Modell werden nun die GI-Variablen hinzugefügt, für welche zuvor gleichfalls monatliche Differenzen gebildet werden. Aus der schrittweisen Modelselektion geht ein erweitertes Modell mit dem GI „Arbeitsmarktservice“ hervor. Im Vergleich zum Benchmark-Modell wird das korrigierte R^2 in der Modellanpassung mit den GI-Variablen von 0,231 auf 0,263 erhöht. Die durchgeführten Hypothesentests führen zu keiner Ablehnung der zugrunde gelegten Modellannahmen.

Tabelle 1: Regressionsergebnisse zur Arbeitslosenrate

	Baseline	Baseline + GI
y_1	-0,548 *** (0,100)	-0,541 *** (0,098)
y_2	-0,239 * (0,099)	-0,243 * (0,097)
GI (ams)		0,012 * (0,005)
constant	-0,0172 (0,040)	-0,015 (0,039)
Adj. R-Squared	0,231	0,263
AIC	94,995	91,943
BIC	105,211	104,713
AICc	95,440	92,617
Breusch-Godfrey (1)	0,181 [0,671]	1,541 [0,215]
Breusch-Godfrey (2)	0,203 [0,904]	2,400 [0,301]
Breusch-Pagan	0,896 [0,639]	1,458 [0,692]
Shapiro-Wilk	0,982 [0,233]	0,982 [0,212]

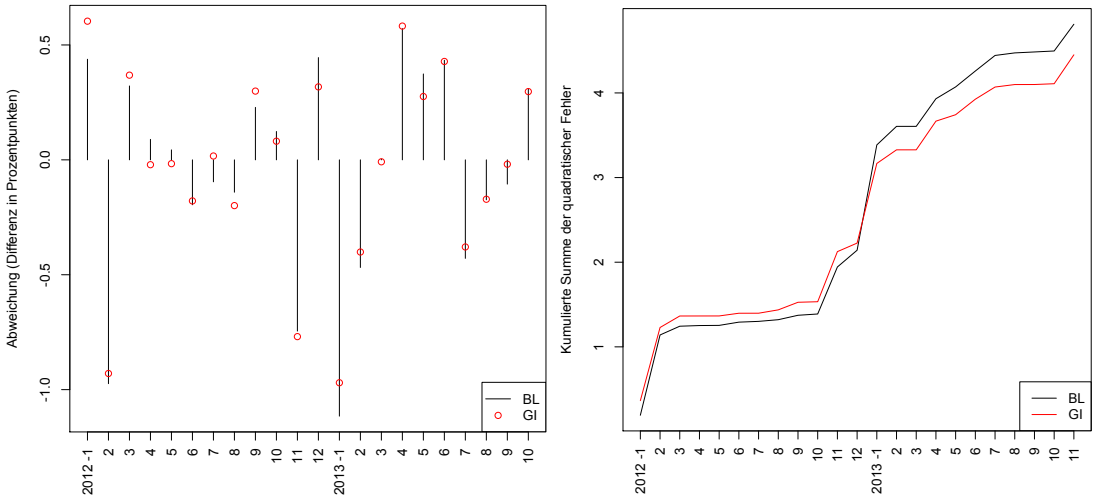
Anm.: Signifkanzcodes: 0 '***' 0,001 '**' 0,01 '*' 0,05 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen auf Basis von Eurostat (2013), Google Trends (2013)

Ein Vergleich der Fehlermaße der Out-of-Sample Nowcasts zeigt eine Verbesserung der Vorhersagen des *Google*-Modells gegenüber dem Baseline-Model. Konkret sinkt der RMSE im erweiterten Modells auf 0,43 (Baseline-Modell: 0,46) und der MAE auf 0,33 (Baseline-Modell 0,36). Das entspricht einer Verminderung des RMSE um 3,8% und des MAE um 5,7%. In Summe sind die absoluten Vorhersagefehler des *Google*-Modells in 68% der Fälle kleiner als die des Benchmark-Modells. Insbesondere in der ersten Periode produziert das Baseline-Modell jedoch bessere Nowcasts, die Vorhersage-genauigkeit des *Google*-Modells liegt erst danach über der des Baseline-Modells. Die Entwicklung der kumulierten Summe der quadratischen Fehler verdeutlich diese Tendenz: Nach den vergleichsweise schlechten Nowcasts zu Beginn des Testsets liegt die Kurve des erweiterten Modells erst ab 2013 unterhalb des Baseline-Modells. In

Anbetracht dessen ergibt sich für das *Google*-Modell trotz des in Summe niedrigeren RMSE und MAE kein eindeutiges Bild bezüglich Signifikanz.

Abbildung 5: Out-of-Sample Nowcasts – Arbeitslosigkeit



	RMSE	MAE	%Better
Baseline-Modell	0,46	0,36	32%
GI-Modell	0,43	0,33	68%

Anm.: Balken stellen die Nowcast-Fehler des Baseline-Modells dar; Punkte die Nowcast-Fehler des erweiterten Modells. Die Ergebnisse des Baseline-Modells sind in schwarz dargestellt, die des Google-Modells in rot.

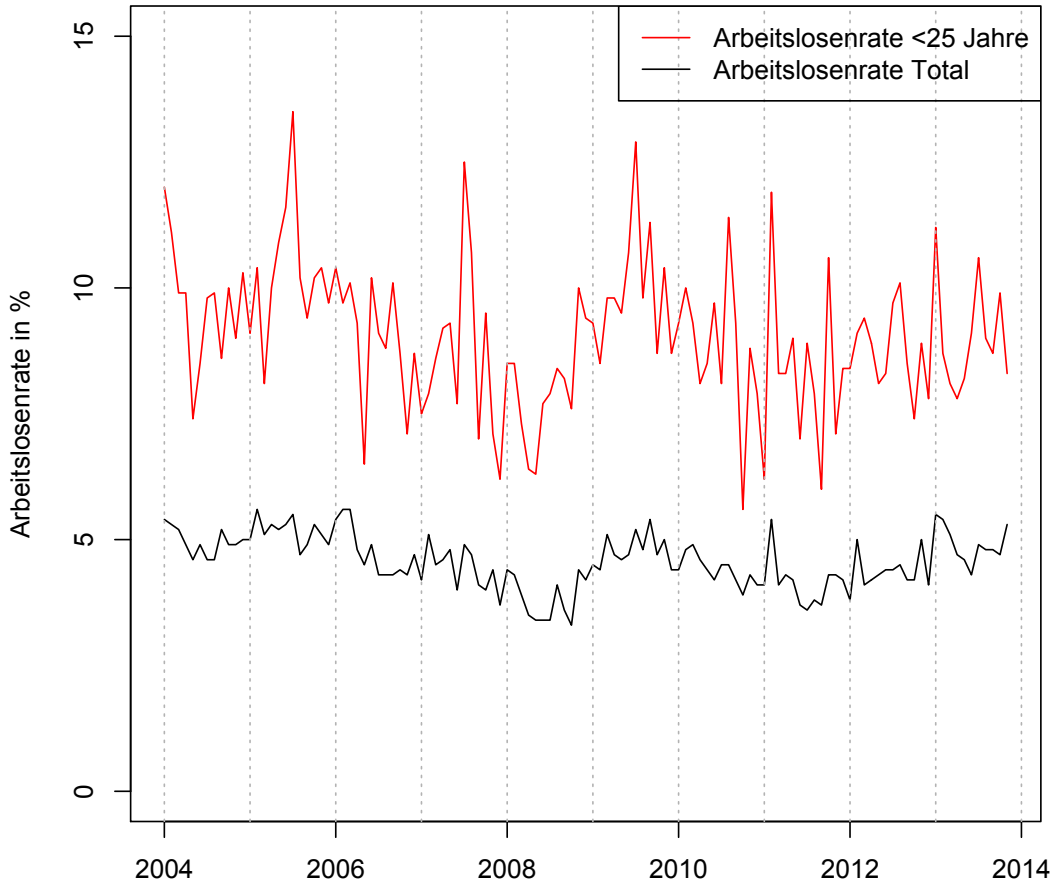
RMSE bezeichnet den Root-Mean-Square-Error, MAE den Mean-Absolute-Error und %Better den Prozentsatz, wie häufig der absolute Prognosefehler des betreffenden Modells unter dem des jeweils anderen liegt.

Quelle: Eigene Berechnungen auf Basis von Eurostat (2013), Google Trends (2013)

5.2 Jugendarbeitslosigkeit in Österreich

Die Arbeitslosenquote für die Kohorte der 15 bis 24-Jährigen, auch als Jugendarbeitslosenquote bezeichnet, liegt in Österreich - wie in anderen Staaten auch - über dem Niveau der Arbeitslosenrate für die übrige Altersgruppe und weist eine stärkere Volatilität auf. Im betrachteten Zeitraum von 2004 bis 2014 bewegt sich die Jugendarbeitslosigkeit zwischen 5,6% im Oktober 2010 und 13,5% im Juli 2005. Aufgrund dieser Unterschiede zur vorhergehend untersuchten gesamtwirtschaftlichen Arbeitslosenrate unterscheidet sich auch das Bild der ACF und PACF (siehe Appendix 4), wenngleich auch diese nicht auf eine Stationarität der Zeitreihe hindeuten. Auch der ADF-Test liegt mit einer Teststatistik von -3,381 (P-Wert: 0,061) außerhalb des 5%-Signifikanzniveaus. Im Folgenden wird auch hier mit den monatlichen Differenzen weiter verfahren.

Abbildung 6: Arbeitslosenquote und Jugendarbeitslosenquote im Vergleich



Quelle: Eigene Abbildung auf Basis von Eurostat (2013)

Für die Jugendarbeitslosigkeit kristallisiert sich ebenfalls ein AR(2)-Modell als Baseline-Modell heraus. (siehe Appendix 5) Unter den Vergleichsmodellen weist dieses Modell den niedrigsten AICc-Wert auf. Weiters zählt es zu den Modellen mit dem höchsten korrigierten R^2 und keiner der ausgeführten Hypothesentests führt zu einer Ablehnung der zugrunde gelegten Modellannahmen.

Die Erweiterung des Modells mittels der schrittweisen Modellselektion fügt zusätzlich die beiden GI-Variablen „Aktive Suche“ und „Bewerbung“ hinzu. Das korrigierte Bestimmtheitsmaß (adj. R-Squared) des erweiterten Modells liegt über dem des Baseline-Modells und auch nach den Informationskriterien – mit Ausnahme des BIC – ist das erweiterte Modell zu favorisieren.

Tabelle 2: Regressionsergebnisse zur Jugendarbeitslosenquote

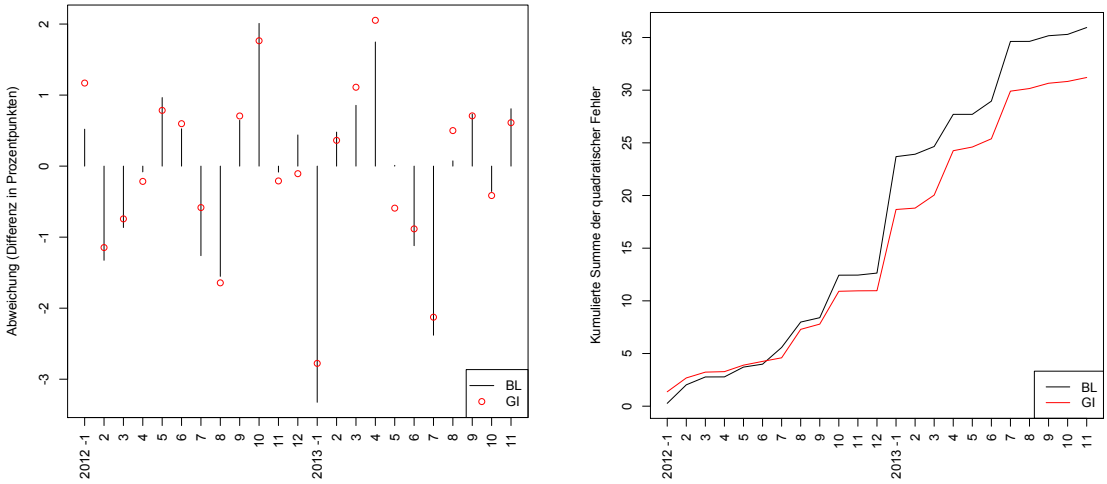
	Baseline	Baseline + GI
y_1	-0,716*** (0,094)	-0,745*** (0,093)
y_2	-0,432*** (0,095)	-0,452*** (0,093)
GI (aktive suche)		0,100** (0,037)
GI (bewerbung)		-0,034 . (0,019)
constant	-0,063 (0,152)	-0,068 (0,148)
Adj. R-Squared	0,378	0,413
AIC	349,109	345,382
BIC	359,325	360,706
AICc	349,554	346,337
Breusch-Godfrey (1)	0,260 [0,610]	0,011 [0,917]
Breusch-Godfrey (2)	3,530 [0,171]	0,013 [0,993]
Breusch-Pagan	2,532 [0,282]	4,473 [0,346]
Shapiro-Wilk	0,986 [0,418]	0,991 [0,778]

Anm.: Signifkanzcodes: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen

Die Out-of-Sample Nowcasts zeigen auch hier eine bessere Performance des *Google*-Modells gegenüber dem Baseline-Modell: Der RMSE sinkt um 6,8% gegenüber dem Vergleichsniveau, der MAE um 1,6%. Die Häufigkeit der Vorhersageverbesserungen durch das *Google*-Modell ist hingegen geringer als im ersten Fallbeispiel. Über das Testset betrachtet ist der absolute Prognosefehler des erweiterten Modells in 52% der Fälle kleiner als der des Benchmark-Modells. Obwohl das *Google*-Modell in diesem Beispiel nicht durchgängig bessere Prognosen liefert, übertrifft es das Baseline-Modell bei einzelnen Punkten klar. Diese Beobachtung wird durch den Verlauf der kumulierten Summe der quadrierten Nowcast-Fehler verdeutlicht. Während die beiden Kurven in den ersten Perioden nahe zusammen liegen, ist der Nowcast-Fehler des *Google*-Modells bei einigen Zeitpunkten vor dem Jahreswechsel 2012/2013 markant kleiner und in der Folge ausschlaggebend für den wachsenden Abstand der Kurven.

Abbildung 7: Out-of-Sample Nowcasts - Jugendarbeitslosigkeit



	RMSE	MAE	%Better
Baseline-Modell	1,25	0,96	47%
GI-Modell	1,16	0,95	52%

Anm.: Balken stellen die Nowcast-Fehler des Baseline-Modells dar; Punkte die Nowcast-Fehler des erweiterten Modells. Die Ergebnisse des Baseline-Modells sind in schwarz dargestellt, die des Google-Modells in rot.

RMSE bezeichnet den Root-Mean-Square-Error, MAE den Mean-Absolute-Error und %Better den Prozentsatz, wie häufig der absolute Prognosefehler des betreffenden Modells unter dem des jeweils anderen liegt.

Quelle: Eigene Berechnungen

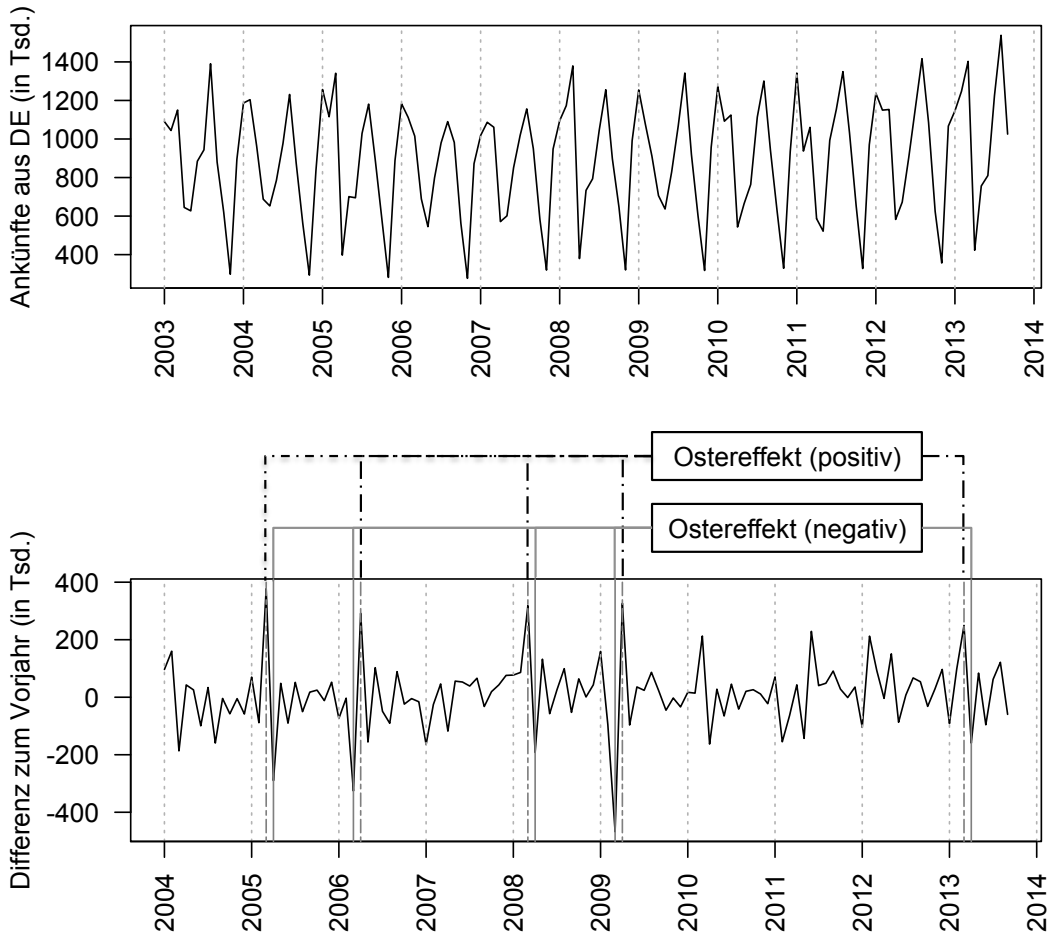
Analog zu den Ergebnissen von Fondeur & Karamé (2013) scheinen sich Daten zum Suchaufkommen der *Google-Suche* auch zur Verbesserung der Nowcast-Genauigkeit bezüglich der Jugendarbeitslosigkeit zu eignen. Grund dafür könnte ein von D'Amuri (2009) angesprochener Selection-Bias sein, der aufgrund der Tatsache entsteht, dass

jüngere Generationen häufiger Internetkanäle bei der Arbeitssuche verwenden. Demzufolge könnten die *Google Trends*-Daten übermäßig stark das Suchverhalten junger Menschen abbilden.

5.3 Tourismus in Österreich - Ankünfte aus Deutschland

Eine Reise nach Österreich unternehmen jährlich bis zu 1,5 Millionen Deutsche. Naturgemäß weist die Entwicklung der Ankünfte eine ausgeprägte Saisonalität auf. In der Wintersaison erreicht die Gästezahl in den Monaten Jänner und Februar den Höchststand, im Sommer in den Monaten Juli und August. Einen weiteren saisonalen Faktor stellen die Osterfeiertage dar, was insbesondere in der Entwicklung der Ankünfte im Vergleich zum Vorjahr zum Ausdruck kommt. Große Differenzen gegenüber dem Vorjahr ergeben sich insbesondere in Jahren, in welchen die Osterfeiertage in einem unterschiedlichen Monat als im vorangegangenen Jahr liegen (in weiterer Folge als „Ostereffekt“ bezeichnet).

Abbildung 8: Ankünfte aus Deutschland (Alle Unterkunftsarten)



Quelle: Eigene Abbildung auf Basis von TourMIS (2013)

Um das saisonale Muster der Besucherankünfte zu berücksichtigen, werden die jährlichen Differenzen der logarithmierten Ankunftszahlen gebildet. Der ADF-Test (Teststatistik = -5,248; p-Wert < 0,01) verwirft für diese Zeitreihe die Nullhypothese einer Einheitswurzel zum 5%-Signifikanzniveau. In der ACF beziehungsweise PACF (siehe

Appendix 6) legen einzelne Spikes, die außerhalb der Konfidenzbänder reichen, eine komplexere Wahl des Modells nahe. Ursächlich für diese lokalen Ausreißer könnten die zeitlichen Verschiebungen des Osterdatums sein. Um die Ostereffekte in den Regressionen zu berücksichtigen werden diese mittels Indikator-Variablen modelliert. Dazu wird zunächst eine Indikatorvariable erstellt, die im Monat des Ostersonntages mit 1 und sonst 0 kodiert ist. Von dieser Indikatorvariable werden jährliche Differenzen gebildet, sodass die Variable den Wert 1 beziehungsweise -1 in Monaten annimmt, deren Ostersonntag in einem unterschiedlichen Monat im Vergleich zum Vorjahr liegt. Diese Indikatorvariable wird als „ostern“ bezeichnet.

Auf Basis des Trainingssets (2004-2011) wird aus verschiedenen Kandidaten-Modellen ein Baseline-Modell bestimmt (siehe Appendix 7). Modell 4 mit den Koeffizienten y_1 und y_12 sowie der Dummy-Variable für Ostern stellt sich dafür als am besten geeignet heraus. Es hat unter den Vergleichsmodellen die relativ höchste Erklärungskraft (gemessen am korrigierten R^2) und ist nach den Informationskriterien (AIC, AICc, BIC) zu favorisieren. Darüber hinaus legen weitere Hypothesentests betreffend der üblichen Regressionsanahmen keine Ablehnung der getroffene Modellwahl nahe.

Im nächsten Schritt wird das Baseline-Modell durch den schrittweisen Modellselektionsalgorithmus mit GI-Variablen erweitert, für welche zuvor gleichfalls jährlichen Differenzen gebildet wurden. Eine Liste der berücksichtigten *Google Trends*-Zeitreihen findet sich unter Appendix 8. Aus der Modellselektion geht ein erweitertes Modell mit den *Google*-Indikatoren „Skigebiete“ und „Länder_Städte“ hervor, wobei für beide die *Google Trends*-Daten aus der Subkategorie „Wetter“ stammen. Die Variable „Skigebiete“ kombiniert Suchbegriffe nach den wichtigsten österreichischen Skigebieten und die Variable „Länder_Städte“ umfasst die Namen der Bundesländer und Landeshauptstädte Österreichs ab. Das korrigierte R^2 dieser Modellanpassung ist mit 0,816 höher als der Wert des Baseline-Modells (0,793) und die Hypothesentests legen keine Ablehnung der zugrundeliegenden Modellannahmen nahe.

Tabelle 3: Regressionsergebnisse zu Ankünften aus Deutschland

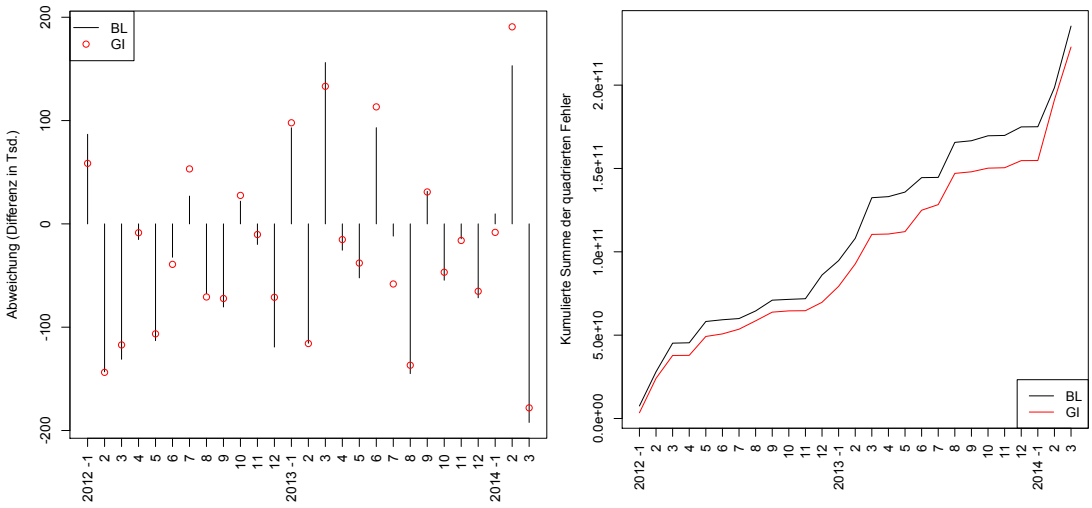
	Baseline	Baseline + GI
y_1	-0,299*** (0,054)	-0,311*** (0,051)
y_12	-0,238*** (0,056)	-0,228*** (0,053)
GI (Nachrichten→Wetter: Skigebiete)		-0,004*** (0,001)
GI (Nachrichten→Wetter:Länder_Städte)		0,003** (0,001)
ostern	0,336*** (0,029)	0,317*** (0,029)
constant	0,010 (0,008)	0,018* (0,008)
Adj.R-Squared	0,793	0,816
AIC	-189,976	-198,160
BIC	-177,822	-181,145
AICc	-189,206	-196,687
Breusch-Godfrey (1)	3,22 [0,073]	0,413 [0,520]
Breusch-Godfrey (2)	3,589 [0,166]	0,457 [0,796]
Jarque-Bera	0,885 [0,829]	2,367 [0,796]
Shapiro-Wilk	0,989 [0,670]	0,986 [0,517]

Anm.: Signifkanzcodes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen

Die Out-of-Sample Nowcasts zeigen eine bessere Performance des *Google*-Modells gegenüber dem Baseline-Modell. Der absolute Fehler des erweiterten Modells ist in 16 von 27 Fällen (59% der Fälle) niedriger als der des Benchmark-Modells. Ebenso kleiner sind die Fehlermaße RMSE und MAE. Hier liegt der RMSE um rund 2,7% und der MAE um 2,6% unter dem des Vergleichsmodells.

Abbildung 9: Out-of-Sample Nowcasts – Ankünfte aus Deutschland (alle Unterkunftsarten)



	RMSE	MAE	%Better
Baseline-Modell	93.374,34	76.904,59	41%
GI-Modell	90.849,56	74.925,37	59%

Anm.: Balken stellen die Nowcast-Fehler des Baseline-Modells dar, Punkte die Nowcast-Fehler des erweiterten Modells. Die Ergebnisse des Baseline-Modells sind in schwarz dargestellt, die des Google-Modells in rot.

RMSE bezeichnet den Root-Mean-Square-Error, MAE den Mean-Absolute-Error und %Better den Prozentsatz, wie häufig der absolute Prognosefehler des betreffenden Modells unter dem des jeweils anderen liegt.

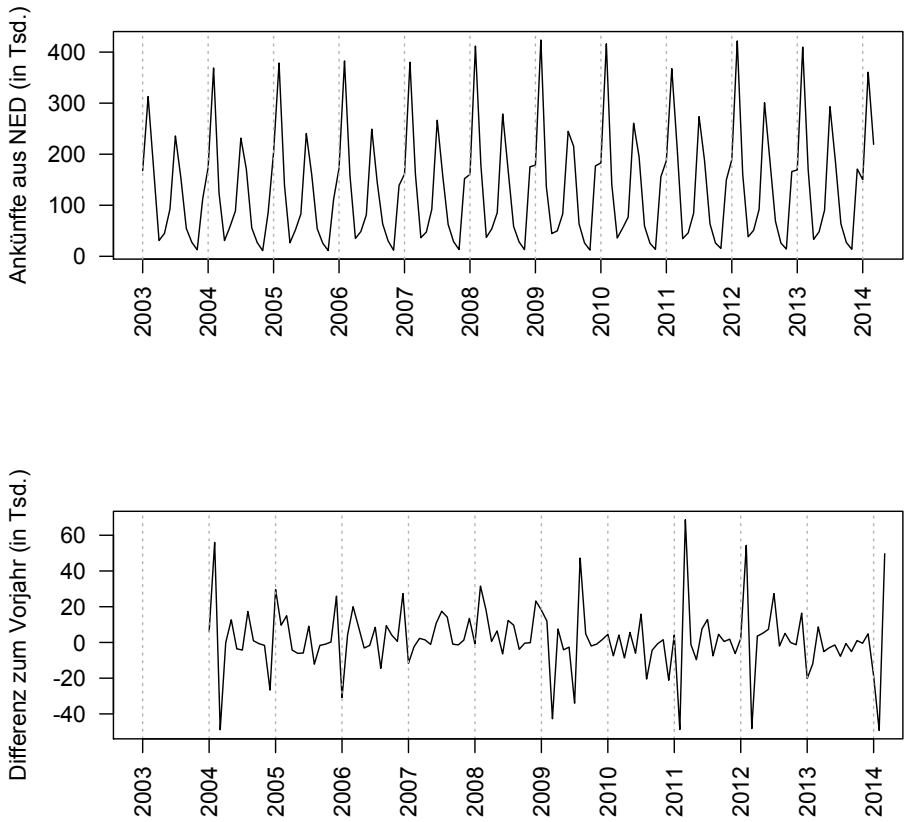
Quelle: Eigene Berechnungen auf Basis von Tourmis (2014), Google Trends (2014)

Obwohl die Nowcast-Fehler des erweiterten Modells zum Teil deutlich über dem Baseline-Modell liegen, lässt sich die Verbesserung des RMSE aufgrund des Verlaufs der kumulierten Summe der quadrierten Fehler doch klarer einordnen. Die Kurve des *Google*-Modells liegt konstant unterhalb der Kurve des erweiterten Modells und produziert demnach in Summe genauere Nowcasts als das Baseline-Modell.

5.4 Tourismus in Österreich - Ankünfte aus Niederlande

Die Niederlande sind mit zuletzt über 1,6 Millionen Besuchern der zweitwichtigste Herkunftsmarkt ausländischer Besucher in Österreich. Im Jahr 2013 besuchten in der Sommer-Saison über 700 Tausend Niederländer Österreich und im Winter waren es über 950 Tausend Gäste. Spitzenmonate sind dabei Februar und Juli mit rund 400 Tausend beziehungsweise 300 Tausend Ankünften.

Abbildung 10: Ankünfte aus Niederlande (Alle Unterkunftsarten)



Quelle: Eigene Abbildung auf Basis von TourMIS (2014)

Aufgrund der Saisonalität werden in der Folge die jährlichen Differenzen der logarithmierten Ankünfte verwendet.[38] Als Baseline-Modell wird auf Grundlage des AICc ein Modell gewählt, dass die Variablen y_1 (Differenz der logarithmierten Ankünfte zum Vormonat) und y_12 (Differenz der logarithmierten Ankünfte mit Lag 12) berücksichtigt (siehe Appendix 10). Die durchgeführten Hypothesentests führen zu keiner Ablehnung

[38] ACF und PACF sind unter Appendix 9 ersichtlich.

der zugrunde gelegten Modellannahmen. Im Vergleich zu den zuvor durchgeführten Analysen hinsichtlich der Besucherankünfte aus Deutschland ist der niedrigere Wert des korrigierten R^2 dieser Modellanpassung auffällig (DE: 0,79; ND: 0,07). Möglicherweise spielen weitere Effekte - etwa zeitliche Verschiebungen von Urlaubs-/Ferienzeiten - ebenfalls eine Rolle. In der vorliegenden Untersuchung können diese jedoch nicht einbezogen werden.

Dem Baseline-Modell werden nun GI-Variablen aus einer Liste korrespondierender *Google Trends*-Zeitreihen hinzugefügt. Eine Aufstellung der berücksichtigten *Google Trends*-Zeitreihen findet sich unter Appendix 11. Als Ergebnis der schrittweisen Variablenselektion wird der GI für Suchbegriffe bezüglich des Wetters als zusätzliche erklärende Variable gewählt. Diese Variable umfasst Suchanfragen nach dem niederländischen Wort für Wetter in Verbindung mit österreichischen Städtenamen oder Regionen (beispielsweise „weer salzburg“). Das korrigierte Bestimmtheitsmaß des erweiterten Modells liegt mit 0,11 über dem des Baseline-Modells (0,07).

Tabelle 4: Regressionsergebnisse zu Ankünften aus Niederlande

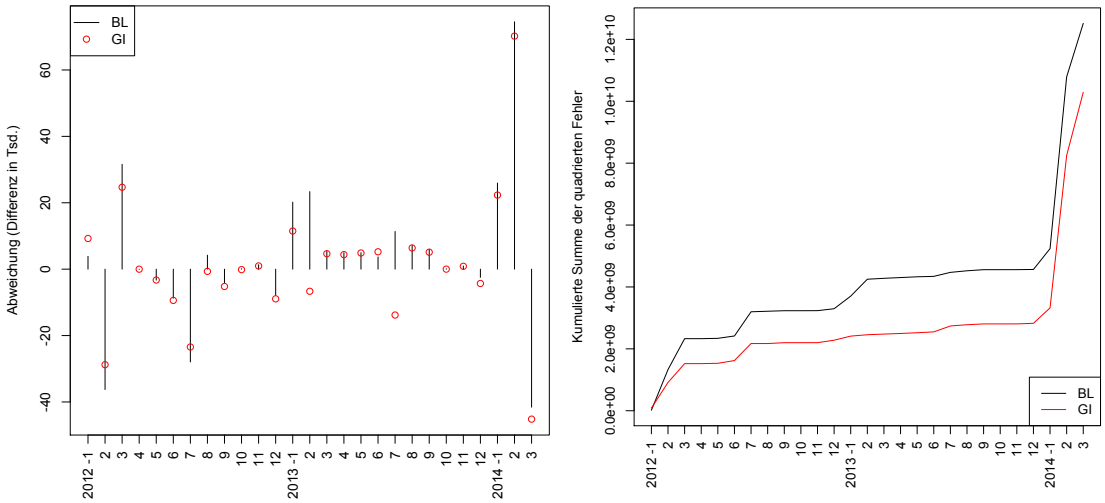
	Baseline	Baseline + GI
y_1	-0,235* (0,102)	-0,253* (0,101)
y_12	-0,193 . (0,102)	-0,216* (0,101)
GI (Wetter)		0,002* (0,001)
constant	0,027* (0,012)	0,018 (0,013)
Adj.R-Squared	0,071	0,107
AIC	-128,101	130,438
BIC	-118,378	-118,284
AICc	-127,595	-129,668
Breusch-Godfrey (1)	0,036 [0,850]	0,254 [0,616]
Breusch-Godfrey (2)	2,895 [0,235]	2,721 [0,257]
Breusch-Pagan	1,521 [0,468]	0,910 [0,823]
Shapiro-Wilk	0,982 [0,281]	0,981 [0,273]

Anm.: Signifkanzcodes: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen auf Basis von Tourmis (2014), Google Trends (2014)

Die Nowcast-Ergebnisse für das Testset zeigen in Summe eine größere Genauigkeit des *Google*-Modells. In 18 von 27 Fällen ist der absolute Vorhersagefehler des erweiterten Modells kleiner als der des Basismodells und der RMSE ist mit 19,524 um rund 9,3% kleiner als der des Vergleichsmodells. Auch nach dem MAE ist die Prognosegenauigkeit des GI-Modells höher: Der MAE des *Google*-Modells ist um 11,6% niedriger als der MAE des Benchmark-Modells. Die höhere Vorhersagegenauigkeit unterstreicht der Plot der kumulierten Summe der quadrierten Fehler, nach welchem die Kurve des erweiterten Modells konstant unter der Kurve des Baseline-Modells liegt.

Abbildung 11: Out-of-Sample Nowcasts – Ankünfte aus Niederlande (alle Unterkunftsarten)



	RMSE	MAE	%Better
Baseline-Modell	21.523,06	13.413,94	33%
GI-Modell	19.523,59	11.860,07	66%

Anm.: Balken stellen die Nowcast-Fehler des Baseline-Modells dar, Punkte die Nowcast-Fehler des erweiterten Modells. Die Ergebnisse des Baseline-Modells sind in schwarz dargestellt, die des Google-Modells in rot.

RMSE bezeichnet den Root-Mean-Square-Error, MAE den Mean-Absolute-Error und %Better den Prozentsatz, wie häufig der absolute Prognosefehler des betreffenden Modells unter dem des jeweils anderen liegt.

Quelle: Eigene Berechnungen auf Basis von Tourmis (2014), Google Trends (2014)

Zusammenfassend kann festgestellt werden, dass die Verringerung des RMSE beziehungsweise MAE im Anwendungsbeispiel mit niederländischen Ankunftszahlen stärker ausfällt als mit deutschen Ankünften und in punkto Signifikanz ein eindeutigeres Bild ergibt. Infolge der Berücksichtigung des Ostereffekts ist die Modellanpassung der Ankünfte aus Deutschland besser als im Fall der niederländischen Ankünfte (korrigiertes R^2 im Baseline-Modell für Deutschland 0,79; für Niederlande 0,07). Möglicherweise

decken die *Google Trends*-Daten für die Niederlande ähnliche Effekte ab, die aufgrund von zeitlichen Verschiebungen von Ferien- und Urlaubszeiten entstehen, und sind demzufolge mit Ursache für die vergleichsweise starke Reduzierung der Prognosefehler. Grundsätzlich unterstreichen die Ergebnisse jedoch das Potential dieser Datenbasis für Anwendungen im Tourismusbereich.

6 Conclusio

Das Internet hat in den letzten Jahren zunehmend Einzug in den gesellschaftlichen Alltag gefunden. Viele Aktivitäten nehmen ihren Ausgangspunkt im Internet oder können überhaupt online abgewickelt werden. Eine zentrale Rolle dabei haben Suchmaschinen, welche als Knotenpunkte den Datenverkehr im Internet registrieren. Im Vordergrund dieser Arbeit steht die Hypothese, dass Informationen über das Suchverhalten von Internetusern eine genauere Bestimmung/Vorhersage des gegenwärtigen ökonomischen Status ermöglichen. Diese Möglichkeit eröffnet sich aufgrund der Tatsache, dass Suchmaschinenstatistiken beinahe in Echtzeit verfügbar sind, während makroökonomische Daten meist mit größerer, zeitlicher Verzögerung veröffentlicht werden.

Die Hypothese wird anhand von je zwei Anwendungsbeispielen aus den Bereichen Arbeitsmarkt und Tourismus untersucht. Als Analysekriterium werden Out-of-Sample-Forecasts von Baseline-Modellen, die Vorhersagen jeweils auf Basis der eigenen Vergangenheit treffen, mit jenen von erweiterten Modellen, die zusätzlich Statistiken zum Suchaufkommen nach bestimmten Begriffen berücksichtigen, verglichen. Die hinzukommenden Daten dafür stammen vom Marktführer unter den Suchmaschinen *Google (https://www.google.com/trends)* und werden zu monatlichen *Google*-Indizes aggregiert.

In Zusammenhang mit den Themen „Arbeitslosigkeit“ oder „Tourismus“ stehen eine Vielzahl an Suchbegriffen, wobei jeder der Suchbegriffe – respektive die Statistiken zum Suchaufkommen dazu - als Prädiktor in den erweiterten Modellen relevant sein könnte. Um aus dem Set möglicher Kandidatenvariablen die bestmöglichen auszuwählen, wird ein schrittweiser Modellselektionsalgorithums angewandt.

Für das Themenfeld „Arbeitslosigkeit“ zeigen die Nowcast-Experimente zur Arbeitslosequote und zur Jugendarbeitslosenquote (Kohorte der 15 bis 24-Jährigen) eine Verbesserung der Vorhersagen bezüglich des RMSE und des MAE. Dazu wurden korrespondierende Suchbegriffe in sechs Kategorien verknüpft und die Baseline-Modelle durch schrittweise Modellselektion mit *Google*-Indizes erweitert. Hinsichtlich der Arbeitslosenquote geht daraus ein Modell mit dem *Google*-Index betreffend der Begriffe „Arbeitsmarktservice“ und „AMS“ hervor, bei der Jugendarbeitslosigkeit sind das *Google*-Indizes der Kategorien „Aktive Suche“ und „Bewerbung“. Die Vorhersageverbesserungen gegenüber den entsprechenden Baseline-Modellen betragen bei der Arbeitslosenquote 3,8% (RMSE) beziehungsweise 5,7% (MAE) und bei der Jugendarbeitslosenquote 6,8% (RMSE) beziehungsweise 1,6% (MAE).

Auch im Tourismus kann die gegenwärtige Entwicklung der Gästezahlen nach unterschiedlichen Herkunftsmärkten mit Hilfe von Suchmaschinenstatistiken genauer bestimmt werden. In zwei Anwendungsbeispielen wird das anhand der Ankunftszahlen von Gästen aus den beiden wichtigsten ausländischen Herkunftsländern der österreichischen Tourismuswirtschaft – Deutschland und Niederlande – gezeigt. Der RMSE und MAE des erweiterten Modells liegt in beiden Fällen unter dem des Baseline-Modells. Für die Ankunftszahlen deutscher Gäste beträgt die Verbesserung durch die Berücksichtigung von *Google*-Indizes 2,7% (RMSE) beziehungsweise 2,6% (MAE). Bei den Ankunftszahlen niederländischer Gäste fällt die Prognoseverbesserung mit 9,3% (RMSE) beziehungsweise 11,6% (MAE) noch deutlicher aus. Beachtlich ist, dass die Modellselektion für diese Aufgabe ausnahmslos *Google*-Indizes zu Suchbegriffen bezüglich des Wetters wählt. Auch die graphische Analyse der Vorhersagefehler unterstützt die grundlegende Hypothese dieser Studie.

Zusammenfassend zeigen die Resultate durchwegs Verbesserungen in den Nowcast-Experimenten. Dennoch stellen sich weitere Aufgaben für die Zukunft: Die hier angewandte Methodik wurde gewählt, um unterschiedliche Anwendungsgebiete analysieren zu können. In konkreten Aufgabenstellungen bleibt jedoch zu untersuchen, ob mit Suchmaschinendaten erweiterte Modelle auch gegenüber State of the Art-Modellen Verbesserungen erzielen können. Zu beachtende Aspekte dabei sind die Berücksichtigung der unterschiedlichen Periodizität der Zeitreihen, die Art der Modellierung sowie die Methodik der Variablenselektion. Summa summarum demonstriert diese Arbeit aber das Potential dieser Datenbasis in den Anwendungsgebieten.

7 Literatur

Anvik, C., & Gjelstad, K. (2010), “Just Google it.” Forecasting Norwegian unemployment figures with web queries. Thesis, Center for Research in Economics and Management, CREAM Publication No. 11-2010.

Artola, C., & Galán, E. (2012). Tracking the future on the web: construction of leading indicators using internet searches. *Banco de Espana Occasional Pape*r, (1203).

Askitas, N., & Zimmermann, K. F. (2009), Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly* (formerly: *Konjunkturpolitik*), Duncker & Humblot, vol. 55(2), Berlin , S. 107-120.

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Nowcasting and the real-time data flow, Working Paper Series, No. 1564, European Central Bank.

Banerjee, A., Dolado, J. J., Galbraith, J. W., & Hendry, D. F. (1993): Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data, Oxford University Press, Oxford.

Barreira, N., Godinho, P., & Melo, P. (2013). Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *NETNOMICS: Economic Research and Electronic Networking*, *14*(3), 129-165.

Bughin, J. R. (2011). 'Nowcasting' the Belgian Economy. Verfügbar unter: SSRN 1903791. http://dx.doi.org/10.2139/ssrn.1903791 [Zitiert am 19. Mai 2014].

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach (2nd ed.), Springer, ISBN 0-387-95364-7.

Chadwick, M. G., & Sengul, G. (2012). Nowcasting unemployment rate in Turkey: Let’s ask Google. Working Papers 1218, Research and Monetary Policy Department, *Central Bank of the Republic of Turkey, Ankara*.

Choi, H., & Varian, H. (2009a), Predicting the Present with Google Trends. Technical report, *Google Inc*. Verfügbar unter: http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf [Zitiert am 25. Jul. 2014].

Choi, H., & Varian, H. (2009b). Predicting initial claims for unemployment benefits. Technical report, *Google Inc*. Verfügbar unter: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/papers/ initialclaimsUS.pdf [Zitiert am 4. Sep. 2013].

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1), 2-9.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. MPRA working paper n. 18403., Verfügbar unter: http://mpra.ub.uni-muenchen.de/18403/ [Zitiert am 17. Mai 2014.].

D'Amuri, F., & Marcucci, J. (2009). 'Google it!' Forecasting the US unemployment rate with a Google job search index (No. 2009-32). Institute for Social and Economic Research.

Fesenmaier, D. R., Cook, S. D., Zach, F., Gretzel, U., & Stienmetz, J., (2009). Travelers' use of the Internet., *Travel Industry Association of America*.

Fondeur, Y., & Karamé, F., (2013). Can Google data help predict French youth unemployment?, *Economic Modelling*, Elsevier, vol. 30(C), pages 117-125.

Gawlik, E., Kabaria, H., & Kaur, S. (2011). Predicting tourism trends with Google Insights.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., & Brilliant, L. (2009), Detecting Influenza Epidemics Using Search Engine Query Data, *Nature*, 1012–14. Verfügbar unter: http://research.google.com/archive/papers/detecting-influenza-epidemics.pdf [Zitiert am 25. Jul. 2014].

Gumprecht, D., Haslinger A., & Kowarik A. (2011), "Austrian LFS Monthly Unemployment Rates.", *Austrian Journal of Statistics*, Volume 40 (2011), Number 4, 297–313

Mayer, M. (2006) Google Official Blog: Yes, we are still all about search. abgerufen am 30. November 2013, von: googleblog.blogspot.co.at/2006/05/yes-we-are-still-all-about-search.html

Helft, M. (2008) Google's New Tool Is Meant for Marketers - NYTimes.com. abgerufen am 30. November 2013, von: www.nytimes.com/2008/08/06/business/media/06adco.html?_r=4&ref=business&

Mohebbi, M. (2011) Mining patterns in search data with Google Correlate. abgerufen am 1. Dezember 2013, von: googleblog.blogspot.co.at/2011/05/mining-patterns-in-search-data-with.html

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). Google correlate whitepaper. *Web document: correlate.googlelabs.com/whitepaper.pdf.* [Zitiert am 1. Dez. 2013].

Österreich Werbung (2013), 'Customer Journey Online - Österreich Urlauber', *Tourismusforschung der Österreich Werbung*

Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, 3(3), 196-210.

Said, S. E., & Dickey, D. A. (1984): Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika* 71, 599–607.

Saidi, N., Scacciavillani, F., & Fahad, A. (2010) Forecasting Tourism in Dubai, Economic Note No. 8, *Dubai International Finance Centre*

Seybert, H, (2012), "Internet Use in Households and by Individuals in 2012." *Eurostat*, Statistics in focus 50/2012, ISSN 1977-0316.

Shumway, R. H., & Stoffer, D. S. (2011) Time Series Analysis and Its Applications with R Examples. (3rd Edition), Springer New York Dordrecht Heidelberg London.

Suhoy, T. (2009), Query indices and a 2008 downturn: Israeli data. Discussion Paper No. 2009.06, Research Department, *Bank of Israel.*

Support.google.com (2007). "Wie werden die Daten ermittelt? - *Google Trends*-Hilfe.". https://support.google.com/trends/answer/92768?hl=de (abgerufen am 27 November 2013).

Terasvirta, T., Tjostheim, D., & Granger, C. W. J., (2010). Modelling Nonlinear Economic Time Series. OUP Catalogue, Oxford University Press, number 9780199587155.

Venables, W. N., & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Vickrey, C. (2007) Google Official Blog: What's hot today. abgerufen am 30. November 2013, von: googleblog.blogspot.co.at/2007/05/whats-hot-today.html

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*(6), 565-578.

Weitzel, T., Eckhardt, A., von Stetten, A., Laumer, S., Kaestner, T.A., & von Westarp, F. (2011) Recruiting Trends 2011 - Eine empirische Untersuchung mit den Top-1.000-Unternehmen aus Deutschland sowie den Top-300-Unternehmen aus den Branchen Finanzdienstleistung, IT und Öffentlicher Dienst, Research Report, Otto-Friedrich-Universität Bamberg und Goethe Universität Frankfurt am Main.

Weitzel, T., Eckhardt, A., Laumer, S., Maier, C., von Stetten, A., & Weinert, C. (2014) Bewerbungspraxis 2014 - Eine empirische Studie mit über 10.000 Stellensuchenden und Karriereinteressierten im Internet*,* Centre of Human Resources Information Systems (CHRIS), Otto-Friedrich-Universität Bamberg und Goethe Universität Frankfurt am Main.

Wöber, K. (2002). Das Internet als Transportmittel touristischer Marktforschungsinformationen am Beispiel von TourMIS. *Tourismus Journal*, *6*(1), 25-48. Verfügbar unter: http://www.tourmis.info/material/tourmis_wp_TXD.pdf [Zitiert am 25. Jul. 2014].

R-Packages

Trapletti, A.,& Hornik, K. (2013). tseries: Time Series Analysis and Computational Finance. R package version 0.10-32.

Venables, W. N., & Ripley, B. D. (2002) MASS: Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Zeileis, A.,& Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL http://CRAN.R-project.org/doc/Rnews/

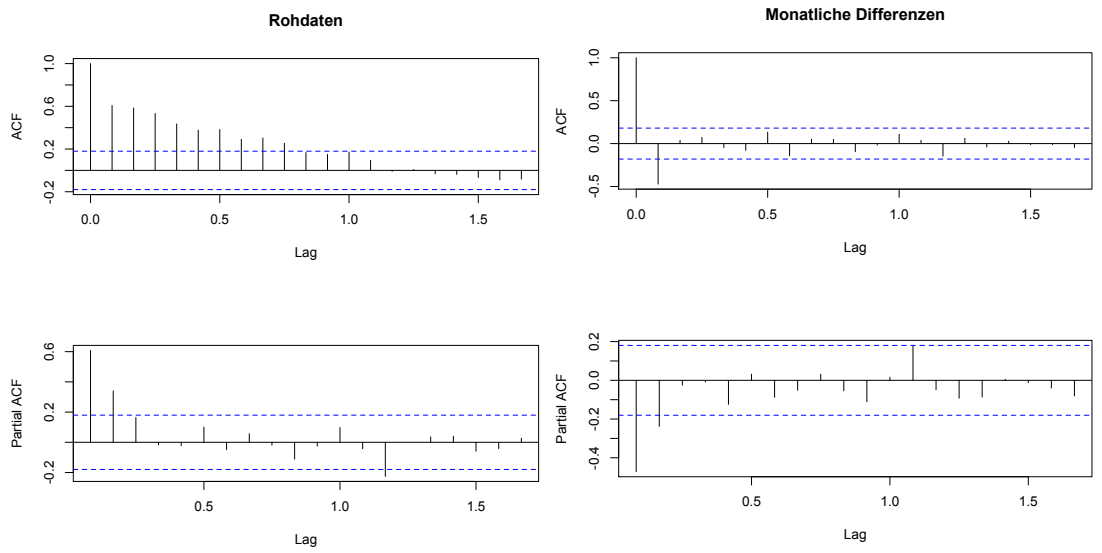
8 Anhang

Appendix 1: Google Trends-Kategorien der ersten Ebene

Autos und Fahrzeuge	Kunst und Unterhaltung
Beruf und Ausbildung	Mensch und Gesellschaft
Bücher und Literatur	Nachrichten
Computer und Elektronik	Naturwissenschaften
Essen und Trinken	Online-Communitys
Finanzen	Referenz
Gesetz und Regierung	Reisen
Gesundheit	Schönheit und Fitness
Haus und Garten	Shopping
Haustiere und wild lebende Tiere	Spiele
Hobbys und Freizeitbeschäftigungen	Sport
Immobilien	Unternehmen und Industrie
Internet und Telekommunikation	

Quelle: Google Trends (www.google.com/trends)

Appendix 2: ACF und PACF der Arbeitslosenrate (Total)



Quelle: Eurostat (2014)

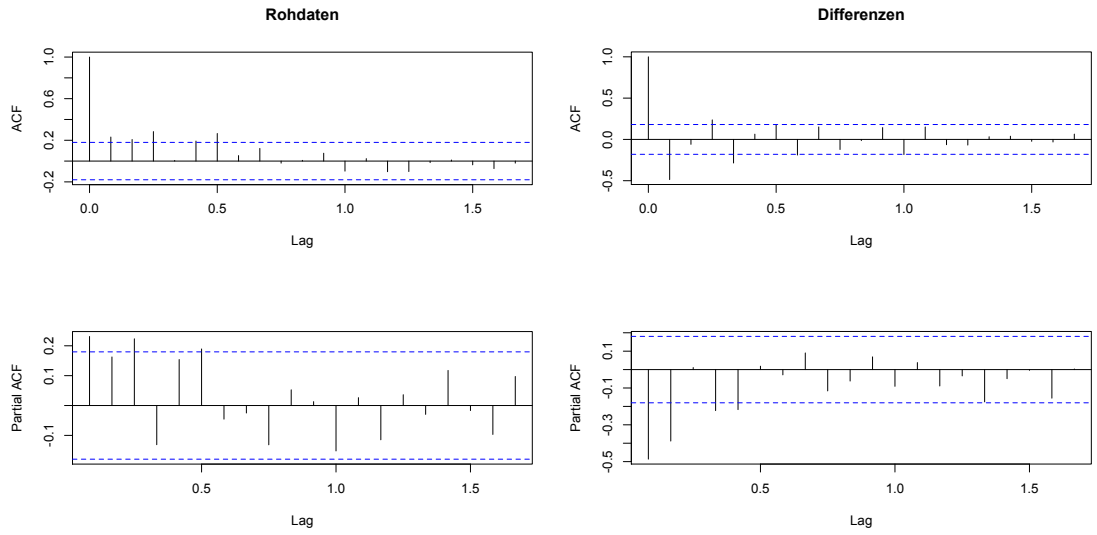
Appendix 3: Regressionsergebnisse der Baseline-Modelle zur Arbeitslosenrate

	1	2	3	4	5
y_1	-0,445*** (0,092)	-0,548*** (0,100)	-0,545*** (0,103)	-0,447*** (0,093)	-0,549*** (0,100)
y_2		-0,239* (0,099)	-0,232* (0,112)		-0,238* (0,099)
y_3			0,014 (0,102)		
y_12				0,042 (0,102)	0,031 (0,099)
constant	-0,016 (0,041)	-0,017 (0,040)	-0,017 (0,040)	-0,016 (0,041)	-0,017 (0,040)
Adj.R-Squared	0,191	0,231	0,223	0,183	0,223
AIC	98,872	94,995	96,975	100,695	96,891
BIC	106,534	105,211	109,745	110,910	109,660
AICc	99,136	95,440	97,649	101,139	97,565
Breusch-Godfrey (1)	6,179 [0,013]	0,181 [0,671]	0,482 [0,488]	6,927 [0,008]	0,368 [0,544]
Breusch-Godfrey (2)	6,182 [0,046]	0,203 [0,904]	0,678 [0,713]	6,983 [0,031]	0,385 [0,825]
Breusch-Pagan	0,408 [0,523]	0,896 [0,639]	1,964 [0,580]	1,044 [0,593]	1,432 [0,698]
Shapiro-Wilk	0,984 [0,301]	0,982 [0,233]	0,983 [0,254]	0,985 [0,348]	0,983 [0,241]

Anm.: Signifkanzcodes: 0 '***' 0,001 '**' 0,01 '*' 0,05 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen

Appendix 4: ACF und PACF zur Jugendarbeitslosenquote



Quelle: Eigene Berechnungen

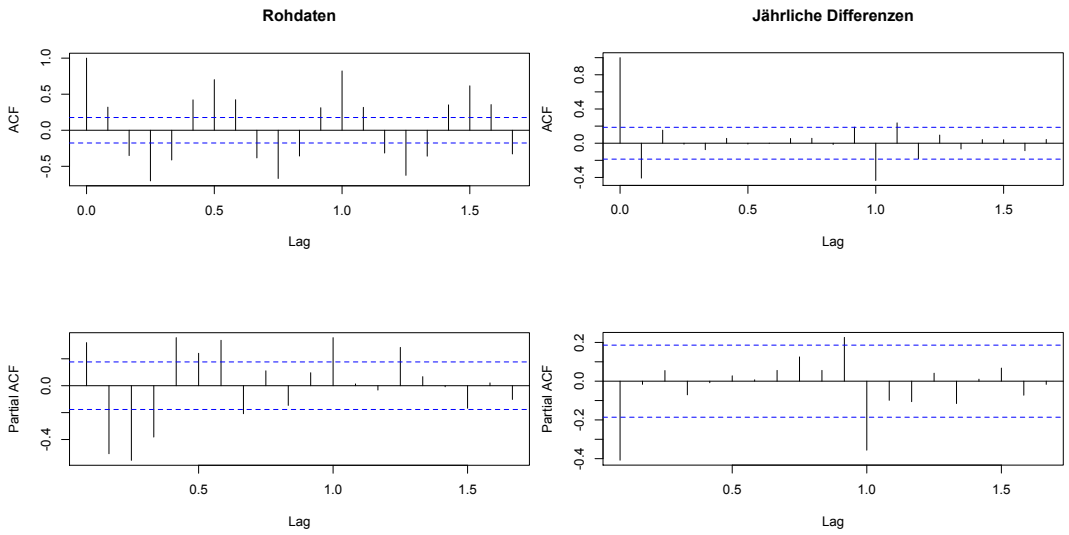
Appendix 5: Regressionsergebnisse der Baseline-Modelle zur Jugendarbeitslosigkeit

	1	2	3	4	5
y_1	-0,505*** (0,090)	-0,716*** (0,094)	-0,690*** (0,104)	-0,483*** (0,091)	-0,694*** (0,095)
y_2		-0,432*** (0,095)	-0,390** (0,119)		-0,428*** (0,095)
y_3			0,064 (0,107)		
y_12				-0,143 (0,107)	-0,126 (0,097)
constant	-0,058 (0,167)	-0,063 (0,152)	-0,061 (0,153)	-0,057 (0,166)	-0,063 (0,151)
Adj.R-Squared	0,246	0,378	0,373	0,253	0,382
AIC	366,267	349,109	350,731	366,438	349,354
BIC	373,929	359,325	363,500	376,654	362,123
AICc	366,531	349,554	351,405	366,883	350,028
Breusch-Godfrey (1)	17,002 [3,733e-05]	0,260 [0,610]	0,492 [0,483]	14,980 [0,0001]	0,355 [0,551]
Breusch-Godfrey (2)	17,300 [0,0001]	3,530 [0,171]	2,584 [0,275]	14,991 [0,0006]	3,123 [0,211]
Breusch-Pagan	6,160 [0,013]	2,532 [0,282]	3,664 [0,300]	6,667 [0,036]	2,920 [0,404]
Shapiro-Wilk	0,980 [0,145]	0,986 [0,418]	0,988 [0,573]	0,982 [0,228]	0,987 [0,474]

Anm.: Signifkanzcodes: 0 '***' 0,001 '**' 0,01 '*' 0,05 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen

Appendix 6: ACF und PACF zu Ankünften aus Deutschland



Quelle: Eigene Berechnungen

Appendix 7: Regressionsergebnisse für die Baseline-Modelle zu Ankünften aus Deutschland

	1	2	3	4	5
y_1	-0,355***	-0,346***	-0,329***	-0,299***	-0,280***
	(0,057)	(0,070)	(0,071)	(0,054)	(0,066)
y_2		0,014	0,075		0,032
		(0,067)	(0,080)		(0,061)
y_3			0,094		
			(0,066)		
y_12				-0,238***	-0,240***
				(0,056)	(0,057)
ostern.diff	0,374***	0,375***	0,379***	0,336***	0,338***
	(0,031)	(0,031)	(0,031)	(0,029)	(0,030)
constant	0,010	0,010	0,009	0,010	0,010
	(0,009)	(0,009)	(0,009)	(0,008)	(0,008)
Adj.R-Squared	0,750	0,747	0,750	0,793	0,791
AIC	-175,027	-173,075	-173,168	-189,976	-188,257
BIC	-165,304	-160,921	-158,583	-177,822	-173,672
AICc	-174,52	-172,306	-172,077	-189,206	-187,166
Breusch-Godfrey (1)	0,697	0,619	0,470	3,22	2,708
	[0,404]	[0,432]	[0,493]	[0,073]	[0,100]
Breusch-Godfrey (2)	1,160	1,628	0,614	3,589	3,693
	[0,560]	[0,443]	[0,736]	[0,166]	[0,158]
Breusch-Pagan	0,173	0,300	1,540	0,885	2,715
	[0,917]	[0,960]	[0,810]	[0,829]	[0,607]
Shapiro-Wilk	0,989 [0,678]	0,989 [0,678]	0,988 [0,620]	0,989 [0,670]	0,989 [0,711]

Anm.: Signifkanzcodes: 0 '***' 0,001 '**' 0,01 '*' 0,05 ( ) std. errors [ ] p-values

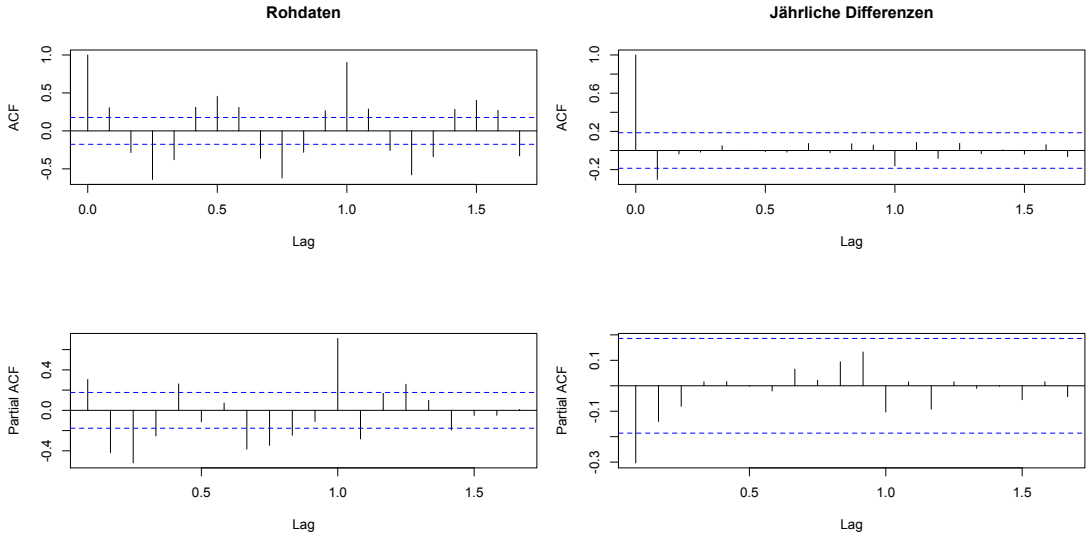
Quelle: Eigene Berechnungen

Appendix 8: Übersicht zu den berücksichtigten Suchbegriffen und Kategorien zu den Besucherankünften aus Deutschland

Bezeichnung	Suchbegriff	in Kategorien
Wetter	wetter österreich + wetter oberösterreich + wetter niederösterreich + wetter vorarlberg + wetter tirol + wetter salzburg + wetter steiermark + wetter wien + wetter kärnten + wetter burgenland	Alle Kategorien Nachrichten→Wetter Reisen Reisen→Reiseziele Reisen→Reiseziele→Berg- und Skigebiete
Regionen	villach + tannheimer tal + graz + kaltenbach + schruns + tschagguns + silvretta + montafon + wilder Kaiser + lienz + salzburg + kitzbühel + gaschurn + partenen + hochjoch + gmunden + mayrhofen + hippach + hohe wand + obertauern + sölden + bad ischl + gallenkirch + zell am see	Alle Kategorien Nachrichten→Wetter Reisen Reise→Hotels und Unterkünfte Reisen→Reiseziele
Skigebiete	ischgl + obertauern + zillertal + sölden + saalbach + hinterglemm + leogang + silvretta + montafon + kitzbühel + serfaus + fiss + ladis + flachau + ski amade + st. anton arlberg + nassfeld + lech + zürs + planai + schladming + warth + schröcken + damüls + obergurgl + hochgurgl	Alle Kategorien Nachrichten→Wetter Reisen Reise→Hotels und Unterkünfte Reisen→Reiseziele
Länder_Städte	österreich + niederösterreich + oberösterreich + burgenland + steiermark + salzburg + kärnten + tirol + vorarlberg + eisenstadt + st.pölten + linz + graz + klagenfurt + innsbruck + bregenz	Alle Kategorien Nachrichten→Wetter Reisen Reise→Hotels und Unterkünfte Reisen→Reiseziele

Anm.: Die Google Trends-Daten wurden ausschließlich für die Region Deuschland geladen.

Appendix 9: ACF und PACF zu Ankünften aus Niederlande



Quelle: Eigene Berechnungen

Appendix 10: Regressionsergebnisse für die Baseline-Modelle zu Ankünften aus Niederlande

	1	2	3	4	5
y_1	-0,223* (0,104)	-0,246* (0,105)	-0,264* (0,105)	-0,235* (0,102)	-0,263* (0,103)
y_2		-0,124 (0,105)	-0,157 (0,105)		-0,148 (0,103)
y_3			-0,169 (0,104)		
y_12				-0,193 , (0,102)	-0,209* (0,102)
constant	0,024 . (0,012)	0,026* (0,012)	0,029* (0,012)	0,027* (0,012)	0,029* (0,012)
Adj.R-Squared	0,042	0,047	0,066	0,071	0,083
AIC	-126,482	-125,938	-126,666	-128,101	-128,226
BIC	-119,190	-116,215	-114,512	-118,378	-116,072
AICc	-126,182	-125,431	-125,897	-127,595	-127,456
Breusch-Godfrey (1)	0,039 [0,843]	0,046 [0,830]	0,974 [0,324]	0,036 [0,850]	0,126 [0,723]
Breusch-Godfrey (2)	2,552 [0,279]	0,096 [0,953]	1,131 [0,568]	2,895 [0,235]	0,290 [0,865]
Breusch-Pagan	0,224 [0,636]	0,182 [0,913]	0,567 [0,904]	1,521 [0,468]	1,474 [0,688]
Shapiro-Wilk	0,983 [0,323]	0,983 [0,330]	0,986 [0,5]	0,982 [0,281]	0,981 [0,246]

Anm.: Signifkanzcodes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1 ( ) std. errors [ ] p-values

Quelle: Eigene Berechnungen

Appendix 11: Übersicht zu den berücksichtigten Suchbegriffen und Kategorien zu den Besucherankünften aus Niederlande

Bezeichnung	Suchbegriff	in Kategorien
Wetter	weer oostenrijk + weer opper-oostenrijk + weer neder-oostenrijk + weer vorarlberg + weer tirol + weer salzburg + weer styria + weer wenen + weer karinthie + weer burgenland + weer bregenz + weer zillertal + weer linz + weer klagenfurt + weer innsbruck	Alle Kategorien Nachrichten→Wetter
Regionen	villach + schruns + tschagguns + silvretta + montafon + wilder kaiser + lienz + salzburg + kitzbühel + gaschurn + partenen + hochjoch + gmunden + mayrhofen + hippach + hohe wand + obertauern + bad ischl + gallenkirch + zell see + wien + ossiach + hermagor + radenthein + kirchberg tirol	Alle Kategorien Nachrichten→Wetter Reisen Reise→Hotels und Unterkünfte Reisen→Reiseziele
Skigebiete	gerlos + ischgl + obertauern + zillertal + sölden + saalbach + hinterglemm + leogang + silvretta + montafon + kitzbühel + serfaus + fiss + ladis + flachau + ski amade + st. anton arlberg + nassfeld + lech + zürs + planai + schladming + warth + schröcken + damüls + obergurgl + hochgurgl	Alle Kategorien Nachrichten→Wetter Reisen Reise→Hotels und Unterkünfte Reisen→Reiseziele
Länder_Städte	oostenrijk + neder oostenrijk + opper oostenrijk + burgenland + styria + salzburg + karinthie + tirol + vorarlberg + eisenstadt + st.poelten + linz + graz + klagenfurt + innsbruck + bregenz + wenen	Nachrichten→Wetter Reisen Reise→Hotels und Unterkünfte Reisen→Reiseziele

Anm.: Die Google Trends-Daten wurden ausschließlich für die Region Niederlande geladen.

Lebenslauf

Matthias Schmidl, Bakk.rer.soc.oec.

Geboren am 2.2.1988 in Wien

Staatsbürgerschaft: Österreich

Kontakt: matthias.schmidl@gmail.com

Ausbildung

seit 10/2013	Magisterstudium Statistik, Universität Wien
seit 03/2011	Magisterstudium Volkswirtschaftslehre, Universität Wien
10/2008 – 03/2011	Bakkalaureatsstudium Volkswirtschaftslehre, Universität Wien Abschluss: Bakk. rer. soc. oec.
2002-2007	HTL Donaustadt, Abteilung für EDV und Organisation Matura am 18.6.2007 mit gutem Erfolg bestanden

Berufliche Erfahrung

seit 11/2012	Wissenschaftlicher Mitarbeiter am Industriewissenschaftlichen Institut (IWI) in Wien
07/2010 – 09/2010	Junior Fellowship am Österreichischen Institut für Wirtschaftsforschung (WIFO)

Sprachen

Deutsch, Englisch

EDV Kenntnisse

Statistiksoftware (R, SPSS Statistics, SPSS Modeller, Stata, EViews)

Ausgezeichnete Kenntnisse in Microsoft Windows, Mac OSX, Linux (Ubuntu)

Microsoft Office (Word, Excell, Access)

Programmiersprachen (Visual Basic, C, C#, SQL)

(wissenschaftliche) Arbeiten

Schmidl, M., & Schratzenstaller, M., (2011): Steuern auf Vermögen und Vermögenserträge: Probleme und Gestaltungsmöglichkeiten für Österreich, in: Wirtschaft und Gesellschaft, Nr. 3/2011.

Dorfmayr, R., Lengauer, S. D., & Schmidl, M., (2013) Struktur und Entwicklung der Industrie Österreichs, in: Industriebuch 2013, Industriewissenschaftliches Institut (IWI), ISBN 978-3-901978-14-2.

Dorfmayr, R., Lengauer, S. D., Lind, T., Luptácik, P., Ramharter, C., Schmidl, M., & Willim, A., (2013) Struktur und Entwicklung der Maschinen & Metallwaren Industrie Niederösterreichs, Industriewissenschaftliches Institut (IWI), Wien.

Brunner, P., & Schmidl, M., (2014) Internationalisierung der Automotiven Unternehmen Österreichs, Industriewissenschaftliches Institut (IWI), Wien.

Brunner, P., Dorfmayr, R., Lengauer, S. D., Schmidl, M., & Schneider, H., (2014) Internationaler Wettbewerb der Wirtschaftsstandorte in der Automotiven Zulieferindustrie – Kurzstudie, Industriewissenschaftliches Institut (IWI), Wien.

Brunner, P., Dorfmayr, R., Schmidl, M., & Schneider, H., (2014) Die Industriestandorte Niederösterreich und Oberösterreich im Vergleich, Industriewissenschaftliches Institut (IWI), Wien.

Dorfmayr, R., Luptácik, P., & Schmidl, M., (2014) Volkswirtschaftliche Effekte der PALFINGER Gruppe in Österreich, Industriewissenschaftliches Institut (IWI), Wien.