SCANNERDATEN – SACHSTAND UND AUSBLICK

Chris-Gabriel Islam, Statistisches Bundesamt

23. Konferenz „Messung der Preise“ am 6. und 7. Juni 2019 in Jena

Übersicht

- Was sind Scannerdaten?
- Was geschah bisher?
- Welche Hürden gibt es momentan?
- Wie geht es weiter?

Was sind Scannerdaten?

- Scannerdaten sind digitale Transaktionsdaten über Umsatz, Menge, Preis und Art der verkauften Artikel auf Ebene der Global Trade Item Number (GTIN), die an den Kassen von Einzelhandelsgeschäften erfasst werden.
- Vorteile gegenüber der traditionellen Preiserhebung:
  - Höhere Genauigkeit durch zunehmende Menge an Beobachtungen
  - Reduzierung des Erhebungsaufwands
  - Zusätzliche Auswertungen möglich, z.B. Bio-Produkte
  - Ausweitung des Erhebungszeitraums
  - Umsatzbasierte Gewichtung möglich

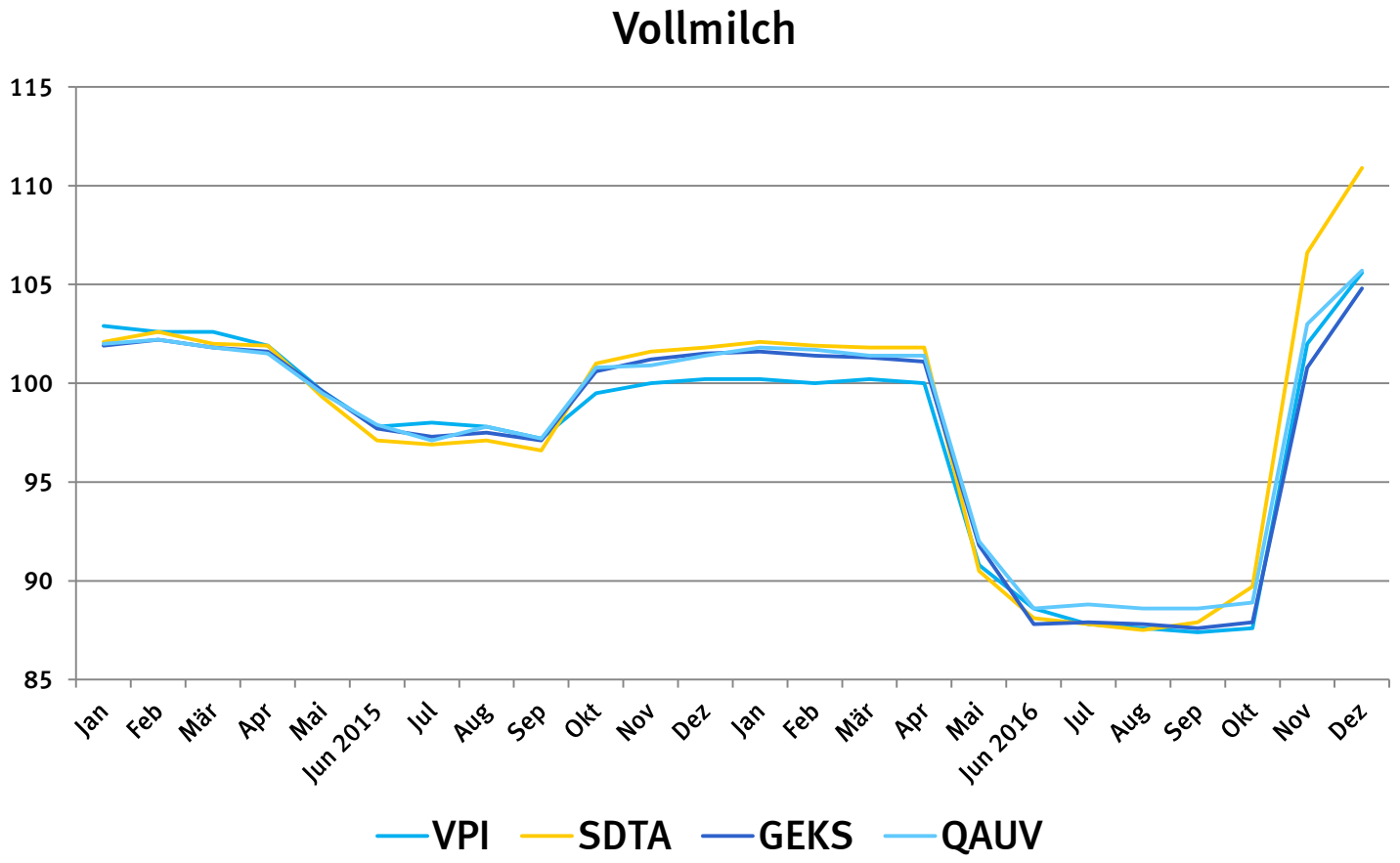
Was geschah bisher?

- Machbarkeitsstudie anhand von Marktforschungsdaten durchgeführt
  - Datenbasis: Scannerdaten des Marktforschungsinstituts Nielsen
    - 44 COICOP-10-Steller aus den Bereichen Lebensmittel und Getränke
    - Wöchentliche Daten von Januar 2015 bis Dezember 2016
  - Auswertung des Datensatzes im Rahmen eines Eurostat-Projektes
    - Untersuchung verschiedener methodischer Fragestellungen
      - Zuordnung der einzelnen Artikel zur COICOP-Klassifikation (Maschinelles Lernen)
      - Prüfung der Stabilität der Produktcodes über die Zeit
      - Testrechnungen diverser Indizes
    - Vergleiche mit den aktuell veröffentlichten Ergebnissen
- Ergebnisse in: Bieg, M. (2019), WISTA 2/2019, S. 25-38.

Was geschah bisher?

GTIN	Produktname	Hersteller	Marke	Umsatz (in €)	Absatz (in Stk)	Inhalt
SENSITIVE	EIGENMARKE WENIG CO2 NORMAL COFF FR 1 1.5 L EW PETFL	HANDEL	EIGENMARKE	15.357,0	3.993	1,5 L
NA	BECK S PILS OA AH 0.33 L RW LONGNECK 2 X 24	BECK'S	BECK S	97,8	2.055	0,33 L
4001686363263	HARIBO BRIXX FRUCHT 200 G BTL	HARIBO	HARIBO	7.104,4	5.126	200 G
4011800262013	SCHWARTAU FRUTTISSIMA HIMBEER 250 G BE	SCHWARTAU	SCHWARTAU	1.451,7	2.982	250 G

Was geschah bisher?



- Daten prinzipiell zur Preisindexberechnung geeignet
- Problem: Daten von Nielsen aufbereitet, gleichen einer Black Box
  ⇨ Originale Scannerdaten notwendig

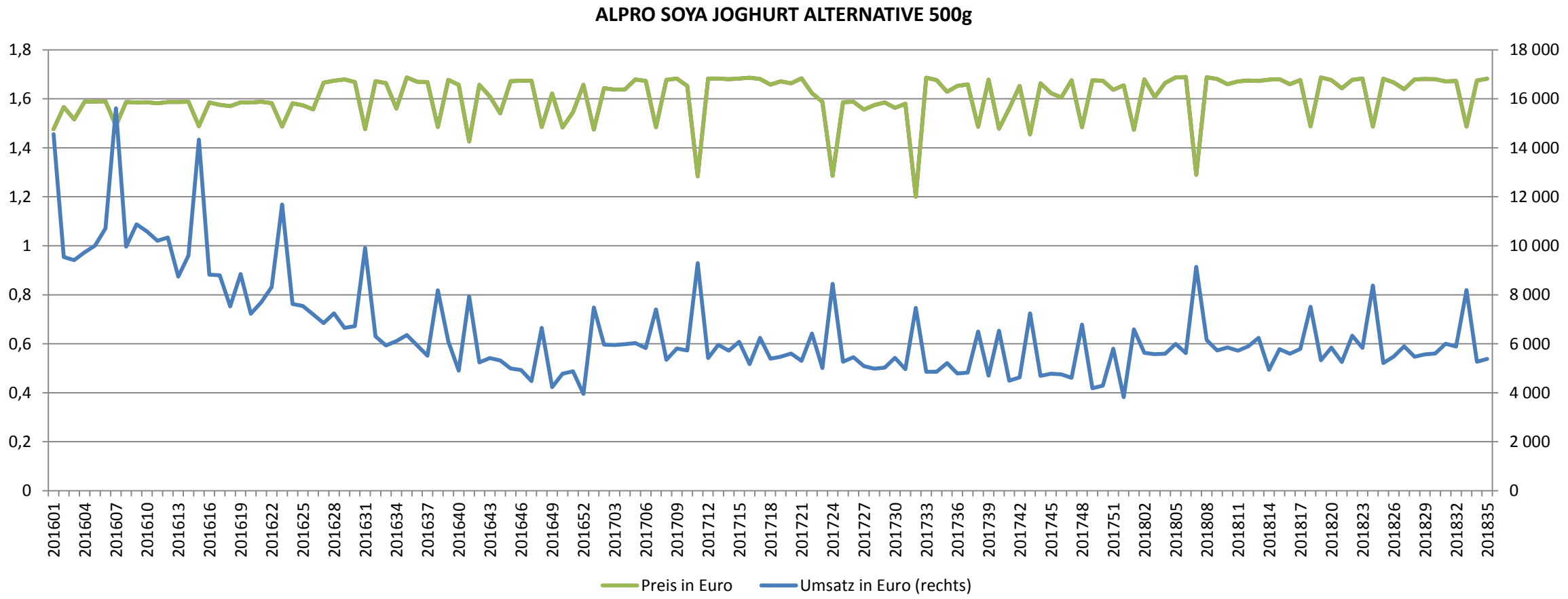
Was geschah bisher?

- Akquise von originalen Scannerdaten
  - Momentane Datenbasis: Scannerdaten eines deutschlandweit aktiven Einzelhändlers
    - Ca. 200 COICOP-10-Steller aus den Bereichen Lebensmittel und Getränke
    - Wöchentliche Daten von Januar 2016 bis KW 19 2019
    - Informationen zu Absatz, Umsatz, Produktmerkmalen, GTIN und Ort
  - Auswertung des Datensatzes im Rahmen eines weiteren Eurostat-Projektes
  - Implementierung eines Prototypen für die Produktion
    - Aufbau einer IT-Infrastruktur zur regelmäßigen Datenlieferung, insb. SFTP-Server
    - Fortführung der methodischen Untersuchungen des ersten Grants
    - Weiterentwicklung der Arbeiten im Bereich des Maschinellen Lernens
    - Festlegen der Berechnungsmethode für die Produktion

Was geschah bisher?

ID	Produktname	Kategorie	Umsatz (in €)	Absatz (in Stk oder KG)	Maß-einheit	Gewicht	Laden-ID
2084966001	WHISKAS TROCKEN 950g	TIERNAHRUNG KATZE TROCKEN BIS 1000 GR.	16,74	6	KG	0,95	8205
4412701001	MERTINGER MILCHHOF SKYR 350g	MOLKEREIPRODUKTE FRUCHTJOGHURT EINZELB. 0.1% BE	1,98	2	KG	0,35	8318
2690081001	NIC.NAPOLEON MEDOC AOC TR., 0,75l	WEIN BORDEAUX	7,49	1	L	0,75	38710
2634941001	QS ORIGINAL INGENHEIMER LEBERWURST CA. 700g 3,5kg/KT	KOCHWURST KRAEUTERLEBERWURST	3,75	0,34	KG	1	8001
2757481001	APFELGRIEBENSCHMALZ 200g	MARGARINE, FETTE, SCHMALZ	-0,06	0	L	0,7	28206

Was geschah bisher?



Was geschah bisher?

- Implementierungsauftrag für einen SFTP-Server abgegeben
  - Bisher liefern sämtliche Einzelhändler Scannerdaten per SFTP an Nielsen
  - Ziel: Lieferweg, der kein striktes Format der Daten vorgibt, um so Einzelhändler in kein Korsett zu zwingen und möglichst viele Daten zu bekommen (Prinzip der Freiwilligkeit)
  - SFTP ist für IT-Dienstleister von Destatis ein Novum, daher konnte erst Anfang Juni mit der Implementierung begonnen werden
  - Deutscher Datenschutz besonders hoch, daher Implementierungsaufwand höher als in anderen Staaten

Was geschah bisher?

- Bund-Länder-Arbeitsgruppe gegründet
  - Teilnehmende Ämter: StLA NRW, HE, BY, BW, ST, TH + Destatis
  - Erstes Treffen: 12.+13.6.2019 in Wiesbaden
  - Ziel der AG:
    - Erfahrungsaustausch
    - Zusammenarbeit mit den Ländern (Preiserhebung gewöhnlich Aufgabe der Länder)
    - Methodischer und technischer Austausch, v.a. hinsichtlich Klassifizierung

Was geschah bisher?

- Nationale Gesetzesgrundlage wird aktuell überarbeitet
  - Europäische Grundlage bereits vorhanden

    Artikel 5 Absatz 4 der Verordnung (EU) 2016/792:
    „Auf Verlangen der nationalen Stellen, … übermitteln die statistischen Einheiten soweit verfügbar elektronische Aufzeichnungen von Transaktionen, z.B. Scannerdaten…“
  - Anpassung der nationalen Gesetzesgrundlage

    Vorschlag: §7b (3) PreisStatG:
    „Zur Erstellung der Statistiken übermitteln die Auskunftspflichtigen den statistischen Ämtern des Bundes und der Länder auf Anforderung elektronische Aufzeichnungen von Transaktionen. Die Aufzeichnungen sind in der Gliederungstiefe zu übermitteln, die für die Erstellung der Statistiken erforderlich ist.“

Was geschah bisher?

» **Produktdefinition auf unterster Ebene**

Figure 1: Traditional Structure

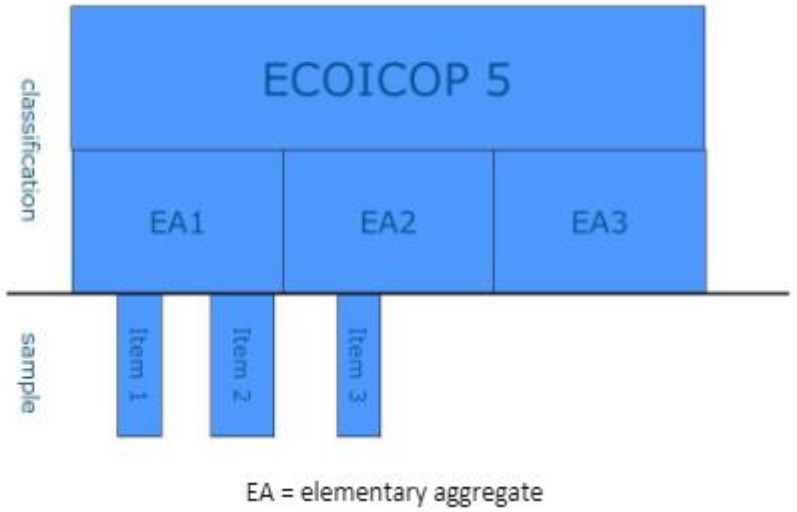


Figure 2: Scanner data

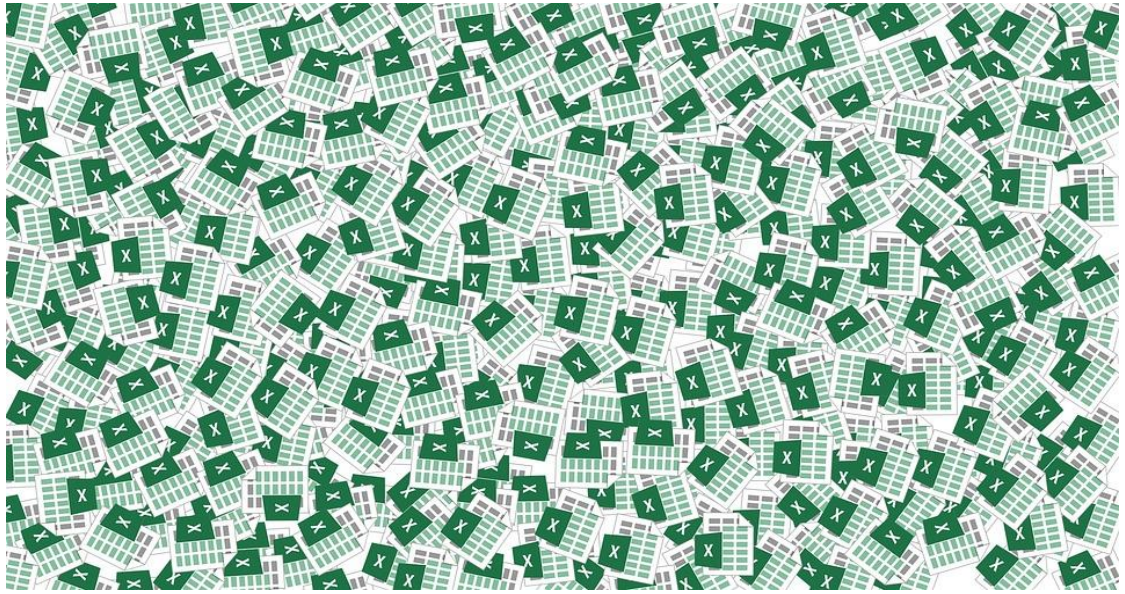
Aus: Eurostat (2017), Practical Guide for Processing Supermarket Scanner Data

Was geschah bisher?

- Produktdefinition auf unterster Ebene
  - Unterhalb des COICOP-10-Stellers werden Durchschnittspreise gebildet, notwendige Voraussetzung: Homogene Produkte, sonst Unit-Value-Bias
  - Problematik bei Scannerdaten größer, da Artikel nicht einzeln ausgewählt werden und in viel größerer Quantität vorliegen
  - Produkthomogenität steht diametral zu Produktkontinuität
  - Chessa (2018) entwarf Match Adjusted R Square (MARS), welcher die Güte einer Produktdefinition als Multiplikation von Homogenität und Kontinuität misst
  - Implementierung bisher in SAS sehr rechenintensiv, besser in R oder Julia?

Was geschah bisher?

- **Produktdefinition auf unterster Ebene**



Es folgt ein Excel-Beispiel

Welche Hürden gibt es momentan?

- **Maschinelles Lernen:**
  - Getestete Algorithmen für originale Scannerdaten auf 10-Steller-Ebene mit Hilfe von Bag-of-Word-Ansatz: Support Vector Machine (77 % Genauigkeit), Random Forest (77 %), Naiver Bayes (81 %)
  - Geplantes Vorgehen:
    - Abschneidegrenze von 90 % erhöht Genauigkeit von Naivem Bayes auf 83 % (10-Steller) bzw. 97 % (4-Steller) bzw. 99,6 % (3-Steller)
    - Manuell 4 % nachzuklassifizieren, entspricht 15 Fälle pro Woche
    - Rest wird mehrstufig klassifiziert
  - Idee: GTIN-Zuordnung anderer (deutschsprachiger) Länder sammeln

Welche Hürden gibt es momentan?

» Maschinelles Lernen:

Artikelname	Kategorisierung des Einzelhändlers	zugeordnete COICOP	COICOP laut Trainingsdatensatz
brombeere, apfel, limette smoothie 250ml	convenience frische saefte frische saefte o&g convenience/salatbar salatbar/eigenproduktion ep frische saefte	Frischkäse	Multivitaminsaft
rq orangen saft 1,25kg	obst zitrus saftorangen obst/gemuese obst zitrusfruechte	Orangen	Orangensaft oder ähnlicher Fruchtsaft
dekoback fondant lila 250g	mehl, baeckereiartikel backartikel dekor-artikel zucker/ mehl/ salz backzutaten/ trockenfruechte backzutaten/ trockenfruechte	Filterpapier, Pappbecher oder Ähnliches	Kaugummi, Gummibärchen oder Ähnliches
rq rahm-kohlrabi 750g	gemuese tk kohlgemuese kohlrabi tiefkuehlkost / eis obst und gemuese tk obst und gemuese tk	Buttergemüse oder anderes tiefgefrorenes Gemüse	Pils, Lager, Schwarzbier o.a. untergäriges Bier
wmf messer	brueherzeugnisse saucen fix-produkte trocken trockenfertiggerichte maggi stammregal maggi aktion	Soßenpulver, Soßenbinder oder Ähnliches	Küchenmesser, Küchenschere oder Ähnliches
frucht secco apfel-joh.himbeer 0,75l ew	saefte fruchtsaft 100 0/0 fruchtsaft 100% ew-flasche alkoholfr.getraenke saefte ew saefte ew	Multivitaminsaft	Weinhaltige Getränke
caffe mauro kaffee gemahlen centopercento 100% arrabica 250g	internationale spezialitaeten kaffee / tee / kakao kaffee / tee / kakao internationale spezialitaeten italienische spezialitaeten italienische spezialitaeten	Salzgebäck	Bohnenkaffee

Welche Hürden gibt es momentan?

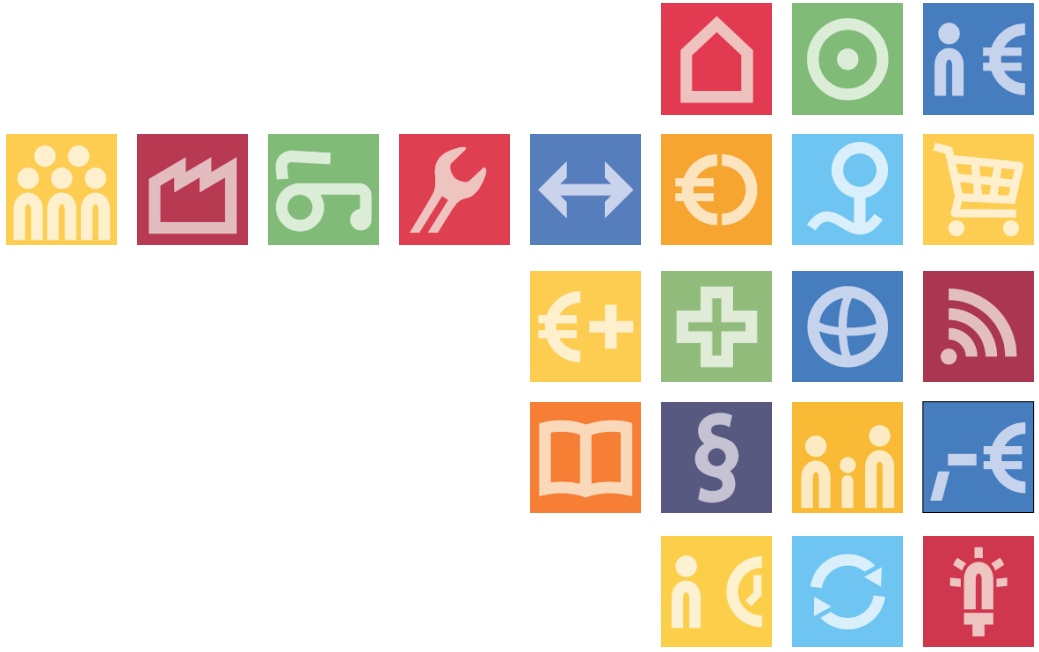
- Aktueller Gesetzesvorschlag reicht zwar für Testzwecke aus, ist aber für eine Produktsetzung zu ungenau (Maxime der Datensparsamkeit)
- IT-Infrastruktur von externem Dienstleister abhängig
- Großer Spielraum, wie Scannerdaten in Index einfließen können, aber keine genauen Vorgaben seitens Eurostat
- Notfallpläne, falls die Datenlieferung eines Einzelhändlers ausfällt (Nielsen?)
- Wie können wir Wiedereinführungen entdecken?
- Umgang mit saisonalen Gütern

Wie geht es weiter?

Meilensteine	Datum
Erstes Treffen BL-AG „Scannerdatennutzung in der Preisstatistik“	12.+13.06.2019
Bereitstellung Dateneingangsserver	15.07.2019
Europäischer Workshop Scannerdata, Thema „Saisonale Güter“	16.+17.09.2019
Regelmäßige Scannerdatenlieferungen von Einzelhändlern gewinnen	2019-2021
Evaluierung von Ergebnissen	Mitte 2020
Parallele Produktion von Indizes basierend auf Scannerdaten im Lebensmittelbereich	ab 2020
Prüfung der Ausweitung von Scannerdaten auf andere Statistiken	offen

» Rollierende Planung je nach Fortschritt bei IT und Datenakquise

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT !



Chris-Gabriel Islam

Telefon: +49/(0) 611 / 75 4832

chris-gabriel.islam@destatis.de

www.destatis.de

BACKUP

1. Stufe: Entwicklung von Methoden und Verfahren

Berechnungen mit Statischem Warenkorb

- Traditioneller Ansatz
  - Dumping Filter
  - Top 70% der umsatzstärksten Produkte je COICOP-10-Steller und Geschäftstyp
  - Fortschreibung bei Produktausfall
  - Ersetzung bei mehrfachen, aufeinanderfolgenden Produktausfällen (länger als 1 Monat)
  - Formel (nach Laspeyres): $I_t^L = \frac{\sum_{i=1}^{N} p_t^i \cdot q_0^i}{\sum_{i=1}^{N} p_0^i \cdot q_0^i}$

1. Stufe: Entwicklung von Methoden und Verfahren

Berechnungen mit Dynamischem Warenkorb

- Quality Adjusted Unit Value (QAUV) Index
  - Kein Dumping Filter
  - Zeitspanne pro Indexberechnung: 13 Monate
  - Art der Verkettung: Movement Splice
  - Formel (nach Chessa): $I_t^{QAUV} = \frac{\sum_{i=1}^{N_t} p_t^i \cdot q_t^i}{\sum_{i=1}^{N_t} v^i \cdot q_t^i} / \frac{\sum_{i=1}^{N_0} p_0^i \cdot q_0^i}{\sum_{i=1}^{N_0} v^i \cdot q_0^i}$

$$\text{mit } v^i = \sum_{z \in T} \varphi_z^i \frac{p_z^i}{I_z^{QAUV}} \text{ und } \varphi_z^i = \frac{q_z^i}{\sum_{s \in T} q_s^i}$$

1. Stufe: Entwicklung von Methoden und Verfahren

Berechnungen mit Dynamischem Warenkorb

- Gini-Eltetö-Köves-Szulc (GEKS) Index
  - Dumping Filter
  - Zeitspanne pro Indexberechnung: 13 Monate
  - Art der Verkettung: Movement Splice
  - Formel (nach Ivancic, Diewert, Fox): $I_t^{GEKS} = \prod_{z \in T} \left(\frac{I_{tz}^{FISH}}{I_{0z}^{FISH}} \right)^{\frac{1}{|T|}}$

$$\text{mit } I_{tz}^{FISH} = \sqrt{I_{tz}^{L} * I_{tz}^{Pa}} \text{ wobei } I_{tz}^{L} = \frac{\sum_{i=1}^{N} p_t^i \cdot q_z^i}{\sum_{i=1}^{N} p_z^i \cdot q_z^i} \text{ und } I_{tz}^{L} = \frac{\sum_{i=1}^{N} p_t^i \cdot q_t^i}{\sum_{i=1}^{N} p_z^i \cdot q_t^i}$$

1. Stufe: Entwicklung von Methoden und Verfahren

Weitere Berechnungen in Planung

- Testrechnungen mit variierenden Parametern
  - Ausreißerbereinigung
  - Dumping Filter
  - Umsatzabdeckung
  - Ersetzungskriterien
  - Zeitspanne pro Indexberechnung
  - Art der Verkettung