물가통계 조사 효율성(빅데이터)과 지수정확성 제고를 위한 연구

2022년 11월

통 계 청
임 성 주

국외훈련 개요

1. 훈련국 : 캐나다

2. 훈련기관명 : The Asia Pacific Foundation of Canada(APFC)

3. 훈련분야 : 통계

4. 훈련기간 : 2020.12.30. ~ 2022.12.29.

훈련기관 개요

명 칭	APF (Asia Pacific Foundation of Canada)
소 재 지	○ 홈페이지 주소 : http://www.asiapacific.ca/ ○ 주소 : 205-375 University Avenue, Toronto ON CANADA M5G 2J5
홈페이지	http://www.asiapacific.ca
설립목적	○ 캐나다와 아시아 관계에 있어 촉진자 역할 및 캐나다에 대한 아시아의 교량 역할을 수행을 위해 1984년 캐나다법률에 의해 설립 ○ 무역, 투자 및 혁신을 통한 아시아와 캐나다의 경제적 관계를 강화하고, 아시아 기후변화, 에너지, 식량안보, 자연자원 관리 등에 대한 해법 제공에 있어 캐나다의 전문성 제공 등을 추진
조 직	○ 이사회, 본부(밴쿠버 소재), 지역사무소(토론토)로 구성 - 본부는 연구부서, 운영 및 행정부서, 홍보부서 등으로 구성
주요기능 및 연구분야	○ 아시아-캐나다 관계에 대한 실행 가능한 연구 및 분석 수행, 정부에 명확하고 구체적이며 실행 가능한 정책 조언과 정보 제공 기능 수행 ○ 캐나다에 대한 아시아국가의 외국인 투자 및 캐나다와 아시아의 무역을 포함하여 아시아와의 관계에 관한 캐나다의 태도에 대한 연례 전국여론 조사 및 설문조사, 통계, 개발, 분석 시행
주요인사 인적사항	○ Marie-Lucie Morin ■ APF의 의장(Interim Chair) ■ 외무부, 국가안보보좌관 등 연방정부에서 30년 경력 ■ 변호사, Université de Sherbrooke 졸업

〈 목 차 〉

Ⅰ 연구 목적 및 연구 내용

1. 연구 목적

조사효율성 측면에서 개인정보 보호 등 조사환경이 악화되고, 새로운 많은 상품 출현으로 현장조사 비용은 증가하고 있다. 따라서, 현장조사를 통한 가격자료수집에서 스캐너 데이터 등 방대한 빅데이터 전환을 통해 조사효율성과 상품의 가격 대표성을 제고할 필요성이 증가하고 있다. 이에, 최근 국제기구(UN, Eurostat, ILO), 주요국을 중심으로 공식통계(물가 등) 편제시 빅데이터 활용에 대한 논의가 활발하게 진행되고 있다.

이미 캐나다, 호주는 물론, 영국, 네덜란드, 프랑스, 이탈리아, 룩셈부르크 등 유럽의 여러 나라들은 이미 스캐너 및 웹스크래핑 자료를 활용해 CPI를 생산 또는 연구를 진행하고 있으며, 자료 수집, 품질 확보, 분류, 품목 정의, 가격 도출, 품질조정 및 지수 산출 전 과정에서 기존 방식과는 다른 소비자물가지수 패러다임의 전환기를 맞이하고 있어 이에 한국 통계청도 능동적으로 준비해 나갈 필요가 있다.

또한 지수 정확성 제고 측면에서는 현행 지수 체계가 지속적인 개선으로 이용자 수요를 충족시켜 왔으나, 신상품, 새로운 업태 출현 등 빠른 경제·사회 구조변화 등 새로운 변화에 직면하고 있다. 생산자물가지수 연쇄지수 전환('13.2월), 광공업생산지수 연쇄지수 전환('18.2월) 등으로 소비자물가지수 체계에 대한 변화 요구도 증대하고 있다

그 동안 통계청에서는 체감 개선 등을 위하여 생활물가, 신선식품, 자가주거비 포함, 농산물 및 석유류 제외, 연쇄방식지수 등 보조지표를 도입하고 정확성과 대표성을 높이기 위하여 기하평균 채택, 조사상품과 대상처 업데이트 등 지속적으로 노력해 왔다. 또한 2013년도에는 소비자 물가지수도 5년 단위 개편에서 2~3년 단위 중간년 개편이 추가되었다. 그러나, 개편 전후 중복기간 시계열 소급 시 지수 수정이 지속적으로 발생, 이용자들의 혼란이 점증되고 있다. 5년 단위 개편뿐만 아니라, 최근 '17년 기준

가중치 적용 후, 기존 '15년 기준 가중치로 공표된 '17.1월~'18.11월 23개월 지수가 수정되어 계약당사자 등 이용자들 애로를 야기하고 있다. 국가통계 주요지표로서의 중요성, 활용성, 고품질통계 요구 증가와 국민들의 체감하는 지표를 공표하도록 지속적으로 요구되고 있는 실정이다.

2. 주요 연구 내용

국내자료 관련해서는 물가지수 계산 및 다양한 산식 자료를 수집하고, 연쇄지수 산식 및 특성, 지수상향 편의 문제, 기존 국내 스캐너 데이터 및 온라인 물가지수 활용방안 관련 연구 자료를 수집 및 검토하였다.

BLS물가지수 conference('21.4.), UNECE CPI 전문가회의('21.6.), Eurostat Workshop Scanner Data and Web Scraping('21.10.), UN Big data Webinar('21.11.), BLS Data 이용자 conference('22.4., '22.5.), OTTAWA 물가지수 전문가 회의('21.6.) 등에 참여하여 최근 소비자 물가지수 관련 논의 과제를 수집・분석하고자 하였다.

스캐너 및 온라인 물가통계 작성 방법 관련, ILO CPI 매뉴얼을 통해 이미 경험을 갖고 있는 일부 유럽 관행에 기초한 IT 시스템 개발, 분류, 지수산식, 품질조정, 법·제도적 측면, 조사자료와 빅데이터 자료 연계방법 등을 검토하였다.

Eurostat 스캐너 및 web scraping, 다변지수 가이드를 통해,

- 스캐너 데이터 관련 개별제품 가격 측정, 여러 지수 산출방법 및 법적 framework, 시뮬레이션 사례를 검토하였다.
- web scraping 관련 법률적 측면, 기술, 품목 적용범위 및 표본, 분류 및 검증, 지수 산출 및 자료 통합 등을 검토하였다.
- 다변지수 관련 데이터 전제조건, 산출 방법, 개별 제품특성, 다변 지수 특성 등을 세부적으로 검토하였고, 양변지수, 가중 및 비가중 다변지수의 상품 시계열 사례별 비교를 통한 적합한 산출방식을 검토하였다.

영국 통계청 사례 등을 통해 CPI 정확성 및 효율성 제고를 위한 스캐너 자료 지수처리 방법, 시간범위, 재출시에 따른 제품식별, 일관성 있는 측정단위, 가격도출, 실제 지불하는 할인 처리 등 개선방안을 검토하였다.

캐나다 CPI지수 작성 및 가중치 개편 방법 등의 다음 사항을 연구하였다,

- 모집단, 조사시기, 가중치 출처, 표본크기, Lowe 등 작성방법
- FAQ, 지수 수정 없는 연쇄지수 도입을 위한 산식 및 가중치 업데이트
- 의류 및 신발지수(웹 스크래핑), 항공운송지수(API) 개선, Covid 영향 반영 및 CPI 품질을 강화하기 위해 국민계정(SNA) HFCE 및 스캐너 등 대체자료 통합을 한 '21년 가중치 개편 자료

또한, 미국 소비자물가지수 주요 작성방법을 다음과 같이 검토하였다.

- 계절조정 방법 및 계절조정 사전조정 개입분석, 표준오차, 지수변화 계산
- 가격조사, 품목 교체 및 품질조정(직접 비교, 직접 품질조정, 대체 방법 등)

본 연구는 우선 Ⅱ장 스캐너 데이터에서는 Eurostat 및 ILO, UN 스캐너 데이터 CPI 활용 가이드 및 매뉴얼을 검토한 결과를 토대로 자료확보 및 법률적 측면, IT 시스템 개발 및 품질보증, 분류 및 개별품목 정의, 여러 지수 산출방법에 대해 검토하였다. 사례로서 영국과 룩셈부르크 통계청 스캐너 데이터 생산 진행과정 그리고 Eurostat와 영국 통계청 산출방법 시뮬레이션 자료를 살펴보았다. 또한 새로운 방법과 경험적 결과에 대한 평가, 사용자 및 이해관계자와의 커뮤니케이션, "새로운" 가격지수의 발표 및 보급에 대한 살펴보고자 하였다. Ⅲ장 웹스크랩핑에서는 Eurostat 및 ILO CPI 매뉴얼을 바탕으로 웹스크래핑 이점 및 한계, 법률적 측면, 웹스크래핑 프로세스, 품목 적용범위 및 표본, 분류 및 데이터 검증, 지수산출 및 자료통합에 대해 검토하였다. 사례로서 캐나다 통계청 의류 및 신발지수, 항공운송지수를 살펴보았다. Ⅳ장 작성방법 관련, 캐나다 작성방법 개요, 지수산식, 가중치 개편 및 지수 연결방법, 계절조정 방법 관련하여 검토하였다. 미국은 작성밥법 개요와 계절조정 방법, 헤도닉을 포함한 품질조정 방법을 살펴보았다. Ⅴ결론 및 제언에서는 앞에서 검토한 내용을 요약하고, 연구를 통해 얻어진 시사점과 이슈가 되는 부분을 적어두었다.

Ⅱ. 스캐너 데이터

1. 소개

우선, 스캐너자료 소개에 앞서, CPI의 전통적인 면접조사 방식, 스캐너 자료, 웹스크래핑 자료간의 차이점은 아래 표[1]와 같이 요약할 수 있다.

특성	대면조사	스캐너	웹스크래핑
자료 획득	수동	자동	자동(웹사이트 변경시 관리 필요)
범위	표본 대상처 및 상품	해당 소매업체 모든 거래	온라인 소매업체 대량 또는 타겟 표본
메타데이터	제품 설명 및 NSO 지정 특성	제품 설명 및 종종 제한된 특성	제품 설명 및 웹사이트에 기술된 특성
수량 자료	없음(proxy 가능성)	판매 수량	없음(proxy 도출 방안 연구 중)
가격 자료	목록(list price)	거래 가격	목록(list price)
자료 이용 시기	NSO 처리 과정 속도	예정된 자료 전송에 따르므로 시차 발생	거의 즉시
시간 범위	수집 시점	자료획득 이전 매일, 매주, 매월 평균 가능	고빈도(시간, 일, 주, 월) 목적에 따라 가능

통계기관이 운영하는 환경이 변화하고 있다. 빅데이터에 접근하고 조사할 수 있는 기회가 제공되고 있다. 바코드 스캐너 기술은 판매 시점 거래에 대한 정보를 포착할 수 있게 했다. 스캐너 데이터는 볼륨이 크며 날짜, 수량 및 수익을 포함한 개별 거래에 대한 제품 정보를 갖고 있고, 통계청에 제공된 자료는 일반적으로 시간(예: 주)에 걸쳐 집계된다. CPI 정확성을 향상시키고 대면 현장 조사 자료를 수집하는 비용을 줄이는 데 사용될 수 있고, 새로운 제품 자료와 같은 다른 측면에서 품질을 개선할 기회를 제공하는 소스이지만, 데이터 분석 및 처리 비용을 증가시킬 수 있다. 이 정보는 동종 제품에 대한 단위 값을 계산하여 CPI를 작성하는 데 사용되는 가격의 정확성, 품목 표본을 개선시키고 그리고 양 또는 수익 정보를 경제적 중요도에 따라 가중치를 매기기 위해 사용될 수 있다.
스캐너 자료는 일반적으로 CPI 범위 전체를 포함하지 않는다. 대부분 이 정보는 임대료, 자동차, 식당 또는 카페를 포함하지 않고 또한, 대형 소매 체인점에서 사용되고 소규모 독립상점에서는 사용되지 않는다. 최근

1) UN Task Team on Scanner Data, Methods and Applications(2021.11.9.)

스캐너 자료는 국가통계기관에서 점점 많이 이용되고 있지만, 사용 전 해결해야 할 과제도 있다. 새로운 데이터 소스의 가격지수를 산출하는 것은 간단하지 않다.

이에 UN(task team on scanner data), Eurostat(HICP task force), 주요 국에서는 스캐너 데이터의 산출에 대한 지침을 더욱 개발하고 국가 간 비교 가능성을 높이기 위해 노력하고 있다.

2. 획득 및 법률적 측면

2.1. 데이터 획득

2.1.1. ILO

공식통계의 작성에서 스캐너 자료 가치는 시간이 지남에 따라 점점 분명해지고 있다. 통계청이 직면한 문제 중 하나는 스캐너 자료를 얻는 것이고, 소매업 또는 제3자 데이터 공급자에게 요청할 수 있다. 이 두 가지 모두 이점과 과제를 가지고 있다. 여러 통계청은 소매업체와 직접 공급을 협상하여 CPI를 산출하는 데 이를 사용했다. 소매업체에서 직접 데이터를 수집하면 아래와 같은 잠재적 이점이 있고, 이는 협상 능력이 수반된다.

- 비용 없이(또는 최소) 데이터셋 확보
- 데이터 세트에 포함된 품목(item)의 범위
- 동질 정보임을 확실하기 위한 품목 집계 수준
- 시간 범위 및 세부 사항(일, 주 또는 월)
- CPI 처리 요구 사항을 충족하는 데이터 공급에 대한 합의된 일정표
- 통계기관 데이터 쿼리에 응답하는 데이터에 정통한 소매업체 담당자

스캐너 데이터 확보와 소매업체와 직접 협상도 난제가 있다. 주된 과제는 데이터 협상은 품목 수준의 매출액과 수량에 대한 정보를 포함하고 있기 때문에 기밀로 간주될 수 있는 것이다. 또 다른 요인은 통계청과 소매업체 간 관계에 관련한 법적 및 제도적 상황이다. 일부 국가에서는 (통계법)에 어떤 자료가 공급되어야 하는지를 규정하는 것이 필요

할 수 있지만, 다른 국가에서는 당사자 간 구두 합의로 충분하다. 스캐너 데이터를 사용하는 국가에서 경험한 바에 따르면 이러한 협상을 완료하는 데 최소 6개월이 걸릴 것으로 예상한다. 협상은 IT 시스템 및 보고 형식에서부터 기밀 문제에 이르기까지 광범위한 주제와 관련이 있다. 통계청과 소매업체 간 체결된 계약은 일반적으로 양해각서(또는 유사한)로 공식화된다. 각 당사자의 역할과 의무를 문서화하고 합의된 일정에 따라 스캐너 데이터의 지속적인 공급을 보장하는 것을 목표로 한다. 소매업체로부터 직접 얻는 대안으로 중개업자나 시장조사업체로부터 자료를 얻는 것이다. 시장조사회사는 데이터를 제공해야 하는 법적 의무가 없지만, 수수료로 통계청이 데이터를 탐색하고 보다 친숙해질 수 있는 좀 더 오래된 스캐너 데이터를 제공할 수 있다. 이 데이터를 더 잘 이해하면 소매업체 또는 시장조사회사와 협상을 시작하기 전에 요구사항이 명확해질 수 있고, 주요 이점은 단일 또는 소수 데이터 제공자와 다양한 제품 집합과 관련된 여러 데이터 공급을 협상할 수 있다.

2.1.2. Eurostat

국가통계작성기관은 지역, 아울렛, 제품 관련 요건을 결정한 이후 소매업체에게 데이터 확보를 위한 목록을 요구해볼 수 있다. 통계기관이 원하는 목록을 확보하던 혹은 타협을 하던, 다음을 염두 해야 한다.

① 데이터 확보는 법적 제도적 여건 및 소매업체와의 관계에 달려있다.
② 소매업체가 스캐너데이터를 제공하도록 협조를 구하는데 오랜 시간이 걸릴 수도 있다. 소매업체와 신뢰관계를 구축하고 강화해야 한다.
③ 시간이 지나면 가용 데이터가 바뀌면서 통계기관 요구 목록 역시 변할 수 있다.

다음 권장 사항은 자료 확보에 기초가 되며, 그 이유도 설명하고자 한다.

① 가능하면 소매업체로부터 직접 스캐너 데이터를 확보하라.

② 상품코드 수준에서 자료를 수집 할 것을 권장한다. 상품코드별로 추출 할 수 있는 데이터로 판매매출 및 수량, 이를 이용해 산출하는

평균거래가격(단가), 포장 내 내용물, 브랜드명이나 제품 명세와 같은 상품 식별을 위한 추가 정보 및 소매업체별 분류 코드가 있다. 데이터에는 해당 기간을 알 수 있는 표시가 있어야 하고, 수량에 대한 정보(조각, 킬로그램, 리터 등)는 품질 조정에 중요하다.

③ 매일 혹은 최대 주간 단위로 데이터를 수집하고 합산할 것을 권유한다. 주 단위로 데이터를 수집하면 규칙적으로 데이터 전송이 이루어지기에, 소매체인에게 쉬운 방법일 수 있다.

④ 스캐너 데이터를 동질적인 아울렛별로 수집 및 합산하거나, 전국 혹은 요건에 맞추어 지역별로 수집 및 합산할 것을 권장한다.

⑤ 스캐너 자료는 기업거래 및 반품과 같은 원칙상 CPI에서 배제해야 하는 거래를 포함할 수도 있으며, 도입 전에 이 부분을 논의할 필요가 있다.

⑥ 스캐너데이터는 할인 가격으로 판매된 상품도 포함한다. 어떤 할인이 어떻게 포함되는지가 명료해야 한다. 실질적으로는 구분이 어려울 수 있다.

⑦ 소매업체에서의 스캐너 자료 추출과 전송 과정을 안전하게 자동화한다.

⑧ 데이터 제공에 대한 세부사항을 공식 합의서에 명시한다. 중요한 자료이기에 구두계약으로 처리해서는 안 된다. 기밀이 높은 데이터이기 때문에 소매업체는 기밀처리 및 사용처에 대한 보장을 원할 것이다.

⑨ 데이터 품질 프레임워크를 사용해야 한다. 스캐너데이터 품질이 요구사항에 맞는지 명시적이고 체계적으로 평가할 수 있도록 품질 보고서를 사용하도록 권장한다.

2.1.3. UN 스캐너 자료 획득 ; 데이터 공급자 연락에서 확보까지

UN Task Team은 교육내용 개발을 위해 데이터 제공자 접촉에서 데이터 획득까지 “결정 트리“로서 가져올 수 있는 경로를 단계별로 기술하였다. 이는 각 노드에 따른 결과 및 활동과 함께 일련의 의사결정 노드를 통해 전개한 것으로 향후 업무 추진시 시사점이 크다.

다음은 첫 번째~4번째 노드를 통한 경로 및 의사 결정 트리[2]이다.

2) Scanner and Web Scraping Eurostat Workshop, Eurostat, 2021.10, From contact to data provider to reception of data, UN Task Team Scanner Data, Kristiina Nieminem, Federico Polidoro)

첫 번째 노드(교점) - 소비재 시장 분석이다.

- 시장별(식료품, 전자제품, 의류, 신발류 등)로 각 시장 내에서 각 소매 유통채널 특성을 파악하고 비중을 추정한다. 대규모 소매유통이 전체 매출액의 50% 이상일 경우 2차 의사결정 노드로 이동한다.

두 번째 노드 - 각 기업의 매출 비중을 추적하여 대규모 소매유통채널의 집중/단편화 수준 분석을 실시한다.

- 3개 초과 기업이 대규모 소매유통매출의 80%를 대표하는 경우 해당 회사들과 접촉하기 위한 중앙 집중 접근(GS1 전국지사, 기업대표협회)을, 3개 이하 기업이 대규모 소매유통매출의 80%를 대표하는 경우 양자 협의를 추진한다. 통계청이 자료를 획득하는 것이 왜 중요한지 설명하고, 유통대기업에도 인플레이션 정보의 질적 향상 측면에서 긍정적이고, 대규모 소매유통에서 발생하는 인플레이션 기여도 정보 제공이 가능함을 강조한다.

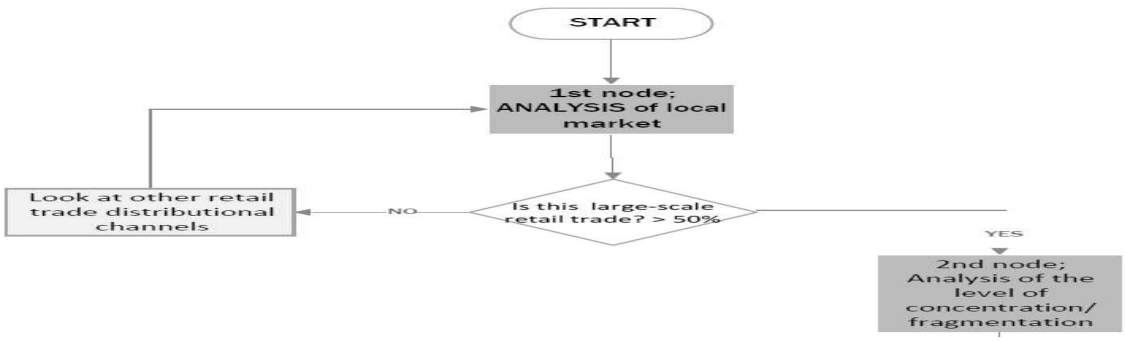
세 번째 노드: 대규모 소매 유통 회사가 통계청에 스캐너 데이터를 제공하는 데 동의합니까, 아니면 꺼려합니까? 그들이 거부하거나 꺼려하는 경우

네 번째 노드: 데이터 수집 부담 때문인가? 만약 그렇다면

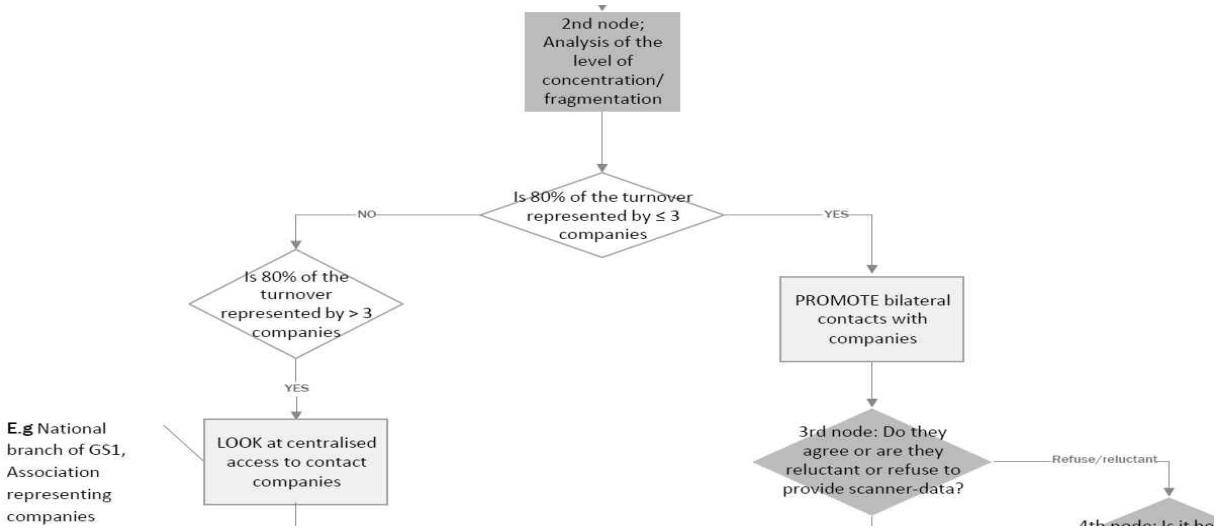
- 데이터 수집・분석업체(닐슨, GfK 등)로부터 자료 확보를 위해 삼자 합의를 추진하다.(통계청은 시장분석업체에서 데이터를 통계청에 보낼 수 있도록 권한을 주고 각 업체에 데이터 제공을 공식적으로 요청) 만약 아니라면, 국가 법률 검토 등이 필요할 것이다.

각 단계별 노드(node)는 다음과 같다.

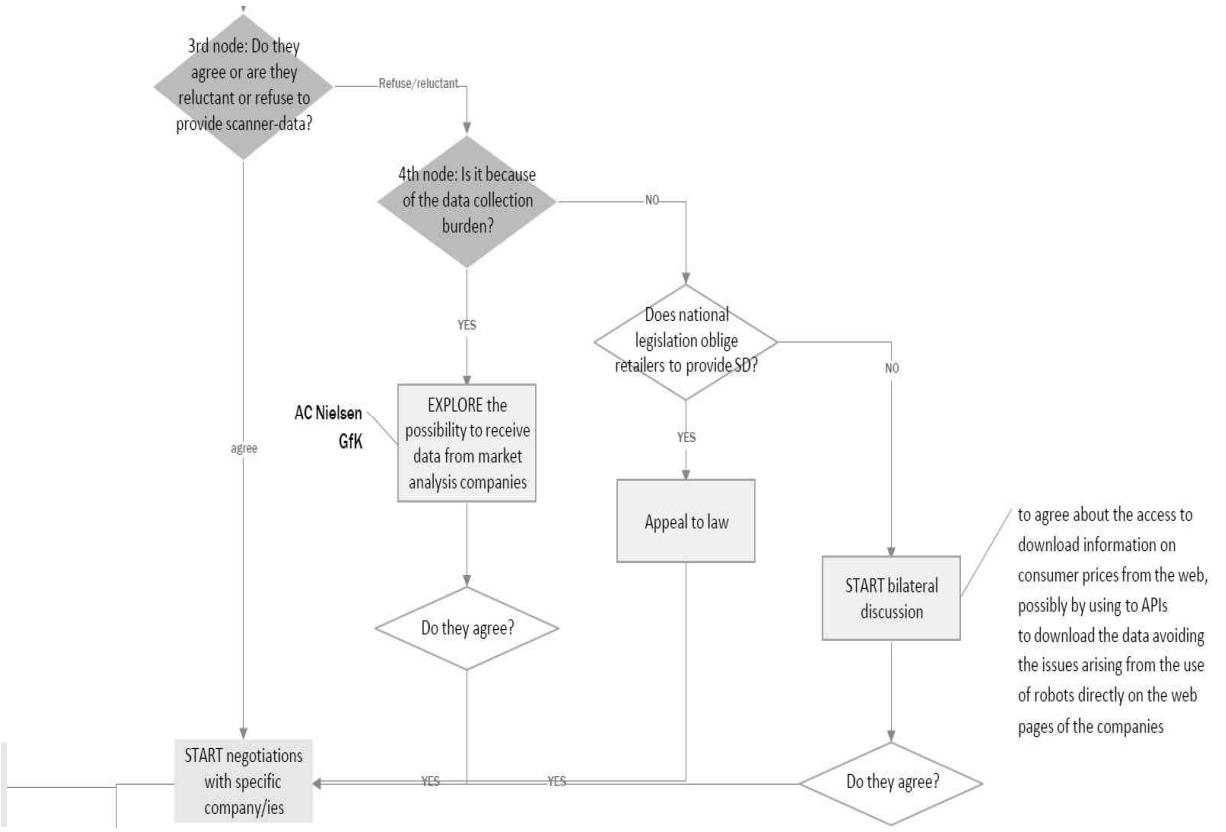
첫 번째 node



두 번째 node



세 번째, 네 번째 node



의사결정 트리와 대체 소스 선택

업체와의 계약 문제 관련하여, 계약 제안서/서명을 통해 업체와의 접촉을 진행하면서 아래와 같은 스캐너 데이터 수신 규칙(적시성 및 발생할 수 있는 문제 또는 실수를 관리하기 위한 연락) 및 데이터 세트 특성에 합의한다.

- 상품설명(Elementary items description, GTIN dictionary)
- 제품 식별(GTIN, SKU, PLU, Prdkey 등을 통해) 및 정보 세분화
- 데이터 내용(주간/일별 데이터, 모든 제품/샘플, 모든 대상처/대상처 샘플, 매출액/판매량/단위의 정의, 거래 유형), 데이터 전송 방식
- 일정, 데이터 전송 빈도 및 전송 형식(csv, 메타데이터, 텍스트)
- 보안, 데이터 보호, 기밀성 및 접근 권한, 답례 서비스
- 데이터 변경시 자료 전송 업데이트 규칙 정의(변경 전 통계청에 통보)

예를 들어, 다음 연도 생산할 물가 지수의 기초를 준비하기 위해 그 이전 1년간의 월별 데이터(세분 자료)의 예비 전달을 요청, 서로 합의되어야 할 데이터 세트의 세분성을 고려하여 사용할 수 있는 다양한 형식의 계약 초안을 작성할 수 있다.

업체가 거부하는 경우 대체 데이터 소스 선택을 고려할 수 있다. 웹에서 소비자 가격 정보 액세스에 대해 합의하기 위해 양자 토론을 진행하고, 가능하면 API를 사용하여 데이터를 다운로드함으로써 회사의 웹 페이지에서 직접 로봇을 사용하여 발생하는 문제를 방지할 수 있다. API접근이 가능하다면 이를 통해 이용 가능한 데이터의 특성과 실행되는 쿼리 규칙(특히 쿼리의 시간 빈도 측면에서)을 해당 소매업체와 의논하여 동의를 구할 수 있다.

2.2. 법률적 측면

스캐너 데이터를 실제 소비자물가지수 작성에 활용하고 있는 국가들의 경우 통계법 관련 근거 또는 소매업체와의 상호 협약을 통해 무료로 제공받고 있다. 안정적인 스캐너 데이터 접근과 제공을 받기 위해서는

관련법 근거 및 소매업체와의 MOU 등 업무협약이 전제조건이라고 할 수 있다. 프랑스 통계청의 경우, 디지털 기술법(Digital Technology Law, 2016.10.7.)은 슈퍼마켓 및 하이퍼마켓 소매점에서 물품(EAN)의 판매와 관련된 모든 데이터(수량, 가격 및 매출)를 민간 소매업자가 정기적으로 제공하도록 규정하고 있다.

3. IT 시스템 개발 및 품질 확보

통계청이 스캐너 데이터를 확보한 경우, CPI를 작성하는 데 효과적이고 효율적으로 사용할 수 있는 정보로 변환해야 한다. 이러한 목표를 달성하기 위해 몇 가지 과제를 가지고 있다. IT 시스템 개발, 데이터 세트의 품질확보 측면에서 살펴보고자 한다.

3.1. IT 시스템 개발

스캐너데이터는 본질적으로 빅데이터이다. 파일 크기는 데이터 특성에 따라 달라진다. 예를 들어, 대상처별 일별 데이터 파일은 소매체인 수준에서 주간 데이터를 집계한 파일보다 용량이 더 크다. 통계청은 CPI 산출하는 데 이를 사용할 경우 대형스캐너 데이터 세트를 획득, 저장 및 처리할 수 있는 IT 시스템이 필요하다. IT 시스템은 분류구조, 형식, 내용이 다른 데이터 세트를 처리할 수 있어야 한다. 소매업체(및 3자 데이터 공급자)가 일반적으로 자체 고유 시스템을 개발하기 때문에, IT 시스템 개발에는 인적 및 재정 자원이 필요하고, 개별 통계청 환경에 따라 다르다. 대규모 투자를 고려할 때, 테스트 데이터(예: 시장조사회사로부터 획득)에 대한 경험을 쌓고 또는 경험이 있는 다른 통계청과 협력하는 것이 중요하다.

3.2. 품질 확보

기존 대상처 방문 가격 수집과 비교했을 때, 스캐너 데이터는 CPI를 산출하는 신 데이터이다. 통계 책임자는 신 데이터 출처가 목적에 적합한 통계 작성 기초를 제공하는지 확인하기 위해 다양한 검사를 수행해

야 한다. 점검은 일상화되어야 하며 각 생산 실행마다 자동으로 수행되어야 하고, 글로벌 검사 또는 세부검사로 나눌 수 있다. 글로벌 검사는 데이터가 생산 프로세스에 진입할 때 수행되며 승인 절차 일부이다. 세부검사는 일반적으로 생산프로세스가 끝날 때까지 수행된다. 글로벌 점검은 통계청이 데이터를 수신할 때 적용되는 광범위한 품질 측정과 관련이 있다. 이 검사를 통해 데이터가 이전 기간에 동일한 데이터 공급자로부터 받은 데이터와 광범위하게 일치하는지 확인할 수 있고, 데이터 세트 형식, 데이터 세트 내 총 품목 수 및 대상처별 수익과 관련될 수 있다. 이러한 글로벌 검사는 데이터 세트의 심각한 오류를 예방한다. 세부검사는 일반적으로 품목 또는 품목군(item group) 수준에서 적용되고 판매 수량, 매출액 및 데이터 세트 내 품목 가격의 중요한 변화를 포착하기 위한 것이다. 가격, 매출액 또는 수량 변동에서 예기치 않은 변화를 점검한다. 스캐너 데이터 처리란 훨씬 더 큰 데이터 집합을 처리하는 것을 의미하며, 기존 수집된 데이터를 처리하는 것과 다른 접근 방식이 필요할 수 있다.

4. 분류 및 개별품목 정의

Eurosat와 ILO CPI 매뉴얼 및 가이드을 검토하였다. 영국에서 논의되는 분류 및 개별 품목 정의에 관해서는 6. 영국 활용사례를 통해 설명하고자 한다. 분류 및 개별품목 정의에 앞서 데이터 전제조건을 살펴보고자 한다.

4.1. 데이터 전제 조건

스캐너 자료를 사용하기 위해서는 몇 가지 데이터 전제 조건을 충족해야 한다. 우선 통계기관이 거래 데이터에 접근할 수 있어야 한다. 데이터 세트 범위와 구조는 데이터 공급자마다 다를 수 있다. 데이터에는 특정 기간 동안 하나 이상의 판매점에서 분류코드(GTIN 또는 SKU 등)별로 매출액과 판매수량이 포함되어야 한다. 둘째, 데이터는 품목을 설명하는 텍스트, 소매업체별 제품 분류코드 또는 제품 특성 등 추가 정보를 포함하는 것이 이상적이다. 셋째, 다변지수를 적용하기 위해서는 지수 산

출기간(즉, time window)와 관련되기에 충분히 긴 데이터가 필요하다.

획득한 자료는 사전 처리 및 분류되어야 한다. 다변지수의 경우 원칙적으로는 모든 거래를 기반으로 하고, 각 제품의 중요도에 따라 통합되기에 제품을 샘플링하거나 매출액이 낮은 제품을 제외할 필요가 없다. 그러나 실제로는 중요한 정보(예: 매출액)가 누락되었기에 또는 다음과 같은 필터 적용을 통해 관측치가 제외될 수 있다.

- 특이치 필터 : 전월(또는 이전 기간)과 비교한 가격 변화가 신뢰할 수 없거나 가격 또는 수량이 비정상적인 경우 관측치가 제외된다. 이는 데이터 세트 코딩 오류, 기타 오류를 가리킬 수 있다.
- 덤핑필터 : 가격과 수량이 모두 이전 기간에 비해 크게 하락할 경우 관찰대상에서 제외될 수 있다. 이것은 염가처분판매에 대한 표시일 수 있다. 이 관측치 들은 어떤 지수 방법을 사용하는지에 따라 결과 지수에 부적절하게 영향을 미칠 수 있기 때문이다.

데이터 분류를 위해 각 관측치는 최소 요구사항으로 하위분류 구조(예 : CPI, HICP, 국가별 분류)에 매핑되어야 한다. 지수 편성에 들어가는 모든 가격 관측치를 분류하는 것이 중요하다.

4.2. 분류

ILO CPI 매뉴얼에서 다루는 스캐너 데이터 분류를 살펴보고자 한다.

스캐너 데이터는 일반적으로 개별 소매점 고유 상품분류를 갖고 있다. 통계청은 CPI 분류와는 다른 상품분류를 갖는 자료를 수신할 가능성이 커서, 이 데이터를 분류하려면 상당한 리소스가 필요하고, 데이터를 처음 수신할 때 가장 큰 규모의 자원 투자가 필요하다. 새로운 상품이 데이터에 진입함에 따라서도 지속적인 분류 리소스가 필요하다. 스캐너 데이터를 CPI 분류로 분류하는 것은 여러 통계청에 의해 다양한 방법으로 설명되었으며, 그 중 다수는 개별소매점 분류를 사용한다. 이러한 분류는 중요한 정보를 제공하며, COICOP의 최저 수준보다 더 상세하거나

같을 경우 매우 유용할 수 있다. 1:1 또는 n:1인 경우(소매업체:COICOP), 스캐너 데이터를 자동으로 매핑할 수 있다. 다른 경우, 데이터는 통계청에 의해 분류되거나 제외된다. IT 시스템과 분류 프로세스는 소매업체 분류의 변화가 적시에 처리될 수 있도록 유연하게 설계되어야 한다. 일부 국가는 시장조사 메타데이터를 구매하여 스캐너 자료를 CPI 분류로 분류했다. 한 통계청은 소매점에서 제공하는 가장 상세한 분류를 사용한 다음 매핑이 올바른지 확인하고 필요한 경우 적절히 변경했다. 일부 통계청은 스캐너 자료 전체 분류를 내부적으로 CPI 분류로 수행했다. 여러 국가는 데이터를 분류하기 위해 머신러닝 알고리듬의 사용을 탐구해왔다. 이 방법은 사전 라벨링된 항목(감독 학습) 또는 라벨링되지 않은 항목(감독되지 않은 학습) 중 하나를 사용하여 각 품목에 대한 올바른 분류법 레이블을 예측하고, 그런 다음 결과 모형을 사용하여 새 데이터 세트를 분류한다. 기계 학습 방법은 소매점에서 사용하는 제품 분류와 CPI 산출에 사용되는 분류 간에 불일치가 있는 경우 특히 유용하다. 이전에 식별되지 않은 새로운 특징을 가진 품목이 적절히 분류되도록 지속적인 유지보수가 필요하다. 스캐너 데이터를 CPI 분류로 분류하는 문제는 자료가 소매업으로부터 직접 확보되었을 때 주로 발생한다. 시장조사회사로부터 자료를 획득한다면 통계청은 CPI 분류에 따라 이미 분류된 데이터 공급을 협상할 수 있다. 일부 국가는 이를 시장조사회사로부터 얻을 수 있는 특별한 장점으로 보고 있다.

다음은 Eurostat 가이드 슈퍼마켓 스캐너 데이터 분류(2017)를 검토하였고, 슈퍼마켓 데이터 활용 관련 자세하게 설명해 주고 있다.

상품코드를 ECOICOP로 분류하는 것은 관련 상품 코드의 양이 방대하기 때문에, 수작업 분류하는 것은 바람직하지 않다. 머신러닝 등 현대적인 기술을 사용하는 게 더욱 적절하다. 분류 방법 연구는 계속적으로 진행되고 있다. 데이터 처리과정에서 첫 번째 단계는 상품을 ECOICOP로 분류하는 것이고, 상품이 매핑되는 분류수준은 통계청이 사용하는 최하위 분류수준(보통 6 또는 7자리)이 될 것이다. 분류는 다음 단계들이 검토되어야 한다.

1) 새 소매업체로부터 스캐너데이터 초기 설정

초기 설정은 상품코드 변동(churn)을 이해하는 심도 있는 연구, 배제해야 하는 상품 및 상품 군 식별, 정기적인 자동화 분류 개발로 이뤄진다. 각 소매업체 마다 다른 초기설정이 필요할 수 있다.

① 분류과정은 가능한 자동화되어야 하고, 자동화 도구를 사용하여 상품 분류를 위한 상품 명세에 대한 분석을 진행 할 것을 권장한다.
② 코드 유형(GTIN, SKU 등), 명세(약어의 의미 등), 코드에 연계되는 메타데이터를 완전히 이해해야 한다.
③ 상품코드 변동(churn)을 이해해야 한다. 이를 위해 장기적인 연구가 필요하다. 코드가 매월 동일한 상품을 가리키는가? 이탈률(attrition rate)은 무엇이며, 새 상품 코드가 도입되는 비율은 어떠한가?
④ 소매업체 분류 사용을 권장한다. ECOICOP의 최하위 단계로 매핑되어야 한다. 소매업체별 분류를 사용해 쉽게 진행할 수 있다. 소매업체가 한 상품을 '백미'로 분류한다면, 그 분류에 들어가는 모든 상품이 ECOICOP 01.1.1.1 (쌀) 하위 범주로 분류될 거라 추정할 수 있다.
⑤ 명세가 불분명한 상품코드 등을 배제하기 위해, 상품 혹은 상품 군이 식별되어야 한다. 가령, 코드는 매일 변하는 '특별 페이스트리 제공' 혹은 '꽃 다발'을 지정하는데도 사용될 수 있기 때문이다.

2) 월별 분류 과정

⑥ 자동 분류 과정을 사용하여 새로운 상품 코드 및 변동하는 메타데이터(명세, 분류 등)를 가진 상품 코드를 매핑한다.
⑦ 명료하게 매핑할 수 없는 상품 코드의 경우, 분류 알고리즘을 이용해야 한다. 불가능 하다면, 새 상품 코드는 수작업으로 분류되어야 한다.
⑧ 다음이 내용을 정기적으로 확인해야 한다.
・다른 그룹으로 옮겨간 상품코드 등 소매업체 분류변화를 모니터링 해야 하고, 상품명세가 바뀐 경우, 해당 상품코드 분류를 확인해야 한다.
・상품코드가 동일한 소매업체별/ ECOICOP 분류를 유지하는지 확인해야 한다.
・불명확한 상품 명세 등을 이유로 배제된 상품코드를 확인해야 한다.

・새 상품 코드와 제거되고 있는 상품코드를 모니터링 해야 한다.(상품 대체를 처리하는 방법의 일환)

⑨ 결과를 재산출할 수 있도록 소매업체별 분류와 ECOICOP로의 분류 이전 버전을 저장해야 한다.

⑩ 분류가 자동화 방식이던 아니던, 그 품질 확인이 필요하다. 무작위로 상품코드 샘플을 선택하여 분류가 정확하게 되었는지 확인할 필요가 있다. 에러는 분류 과정의 개선으로 이어져야 한다.

3) 소매업체별 분류로(with a retailer-specific classification) 분류

소매업체별 분류를 분석한 이후, ECOICOP 하위 단계에 속하거나 일치하는 합산의 모든 상품 코드는 ECOICOP로 직접 연계될 수 있다. 이후 매핑될 수 없는 소매업체별 분류 합산을 분석할 필요가 있는데, 가령 쌀, 야채, 조미료 등을 포함하는 '아시아 식품' 이라는 합산을 분석하는 것이다. 상품코드를 재 할당했는데 상품코드가 적절하게 포함되는 기초 합산의 매출에 상당한 영향을 준다면, 그 상품코드는 할당되어야 한다. 만일 소매업체별 분류합산이 매출 측면에서 중요하지 않다면 뺄 수도 있다. 하지만, 여전히 모니터링을 실시하여 일정 기준을 넘어가는 경우 다시 포함시키도록 해야 한다.

4) 소매업체별 분류가 없거나 사용할 수 없을 때 분류

대안이 필요하다. 수작업으로 상품코드를 분류하기엔 많은 자원이 소요되기에 분류 알고리즘을 사용하는 게 낫다.

4.3. 개별 품목 정의

4.3.1. (ILO) 품목(variety) 정의

지수 계산을 하기 전에 가격을 매길 개별 품목을 정의해야 한다. 기본 원리는 비슷한 것을 비슷한 것을 비교하고 시간 경과에 따라 그 같은

품목 가격을 추적하는 것이다. 일반적으로 제품코드 수준은 데이터에서 가장 상세한 동질성 수준을 나타낸다. 이 제품 차원 외에도 대상처 및 시간 차원도 고려해야 한다. 종종, 동일한 또는 유사한 상점에서 다른 시점에 판매되는 동일한 품목을 고려하여 해당 매장에 대해 평균 거래 가격(단가)을 계산할 수 있도록 충분히 균질한 품목을 정의할 수 있다. 일부 경우에 GTIN이 안정적이고 수명이 길다. 일부 국가는 소매업체 제품코드들, 예를 들어, GTIN보다 더 많이 집계된 SKU코드에 액세스할 수 있다. 이러한 품목 코드 수준은 지수를 계산하기에 너무 상세할 수 있다. 의류와 신발과 같은 일부 제품에서는 품목 코드가 자주 나타났다 사라지면서 시간에 따라 매칭하기가 어려워 가격 변동이 적절히 측정되지 않는다. 상품코드 변경에 대처하기 위한 다양한 전략이 필요하다.

하나의 접근은 유사한 특성을 가진 개별 상품을 묶는 것이다. 품목을 보다 광범위하게 또는 더 좁게 정의할 수 있다. 이러한 그룹 내에서 개별 상품들이 대략 무차별할 수 있도록 그룹을 만드는 것이 중요하다. 이 수준에서 단위 값을 계산하면 품목 간의 대체 효과를 포착할 수 있을 뿐만 아니라 시장에 진입하는 새로운 품목의 포함도 용이해진다. 통계청은 이런 그룹 간 균형을 유지하는 데 있어 절충에 직면해 있다. 그룹핑이 너무 광범위하면 개별 상품이 엄격하게 비교되지 않기 때문에 단위 가치 편향(및 높은 변동성)이 발생할 수 있다. 반면, 그룹을 너무 좁게 정의하면 나가는 상품과 새 상품 또는 반환 상품 간의 매칭이 부족해질 수 있다. 이 단계에서의 결정은 결국 얻어지는 물가 지수에 상당한 영향을 미칠 수 있다. 이는 특히 기술 제품, 특히 회전율이 높은 모델에 관련될 수 있다. 이러한 그룹의 실질적인 구성은 어려울 수 있다. 통계청은 소매점에서 사용하는 내부 분류 코드뿐만 아니라 브랜드와 크기를 포함한 제품 특성에 대한 정보를 필요로 한다. 일부 소매업체는 특정 텍스트 문자열에만 특성을 제공하는 반면, 다른 소매업체들은 상품의 특성을 설명하는 여러 가지 다른 변수를 가질 수 있다. 텍스트 문자열에 수집된 특성을 분류에 사용하기 위해서 어떤 형태의 텍스트 마이닝[3]이 필요할 수 있다. 모든 특성이 똑같이 중요하고 가격에도 같은

3) (출처: 한국정보통신기술협회, 용어사전) 텍스트 데이터에서 가치와 의미가 있는 정보를 찾아내는 기법.

정도의 영향을 미치는 것은 아니다. 개별 품목 군집화는 중요한 가격 결정 특성들에 의해 정의되어야 한다. 또 다른 접근법은 새로운 상품과 소멸상품 가격을 이용할 수 없을 때 대체(impute)하는 것이다. 가격이 헤도닉으로 대체될 수 있고, 상품의 특성을 사용하여 그룹핑을 하는 대신, 현재 결측 가격을 추정하는 데 사용된다.

현장에서 수집한 가격을 스캐너 데이터 가격(단위값)으로 대체하면 일반적으로 자원 절감이 이루어지는데, 이는 CPI 현장 책임자가 더 이상 가격조사 기업을 방문할 필요가 없기 때문이다. 자원절감 가능성은 현장 담당자 감축 규모와 데이터를 관리하고 처리하는 데 필요한 리소스 증가에 영향을 받는다. 단위 값은 품목 구성 변화와 품질 변화가 가격 변동으로 반영되어서는 안 되기 때문에 시간이 지남에 따라 규격이 일정하게 유지되는 단일 동질 품목과 관련되어야 한다. 이러한 요구사항은 현장에서 수집한 가격을 스캐너 데이터 세트의 정보로 대체할 때 몇 가지 문제를 제기한다. 통계청과 데이터 제공자 간 협상은 CPI를 산출하는 데 사용할 단위 값 생산을 지원하는 데 필요한 항목(item) 집계(또는 세분화)의 적절한 수준에서 데이터에 대한 접근을 보장하기 위해 필요하다. 제품 특성의 직접적인 공급은 품목의 분류를 용이하게 할 수 있다. 이러한 정보는 사용 가능한 경우 명시적 품질 조정을 수행하는 데 사용될 수 있다.

여러 NSO는 스캐너 데이터 세트에서 단위 값 데이터를 산출한 경험이 있다. 가장 자세한 수준에서 데이터 세트 품목(item)은 일반적으로 GTIN 또는 그 변형, 범용 제품 코드(Universal Product Code) 등으로 식별된다. GTIN 같은 표준 식별자는 다른 소매점에 걸친 품목 추적이 가능하지만, 소비자와 무관하다고 여겨지는 포장 등 특성별로 구분하는 너무 상세할 수 있다. 그래서 품목 변동(churn)이 과대평가되고 CPI 계산을 방해할 수 있는 재출시의 잠재적 문제가 있다. 예를 들어 GTIN을

많은 정보들이 온라인 뉴스 기사, 기술문서, 도서, 전자 우편 (이메일) 메시지, 마이크로 블로그(micro-blog), 소셜 네트워킹 서비스(SNS) 및 웹페이지와 같은 텍스트 형식으로 저장된다. 이렇게 공개된 다양하고 풍부한 텍스트 정보에서 특정 주제와 관련한 부분을 뽑아 의미를 분석하고 사회 현상이나 여론의 경향 등 고품질의 정보를 도출하기 위한 방법으로 텍스트 마이닝 기법을 활용한다.

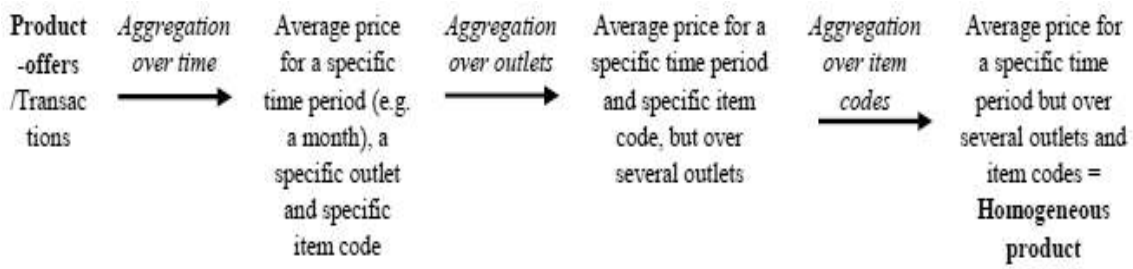
품목 식별자로 사용할 때 GTIN이 변하는 동종 품목의 가격변동은 측정되지 않는다. 가격측정의 필수적 부분은 품질변화와 신규 품목 도입을 고려하는 것이다. 이는 현장조사원이 소매점을 방문하여 연속적 기간에 동일하거나 동등한 품목에 대한 가격 변동을 측정하고 새로운 품목을 식별한다. 품목 특성이 변화됨에 따라 통계청 책임자는 품질변화 효과를 가격변동과 분리해 순수 가격변동만을 측정할 수 있도록 정보를 수집한다. 그러나, 스캐너 데이터는 품질 변화를 고려하는 것은 어렵고 매달 사용할 수 있는 품목에서 높은 수준으로 변동(churn)하는 경향이 있다. 새로운 모델이 출시되고 구형모델은 시장에서 퇴출되고 있다.

4.3.2. (Eurostat) 개별 제품 정의

스캐너 데이터를 취득, 처리 및 분류한 후에는 개별 제품을 구체적으로 명시해야 한다. 개별 제품은 시간이 지남에 따라 가격이 추적되는 통계 단위이다. 개별 제품 가격은 지수 산출의 입력이다. 먼저 개별 제품 개념, 제품을 구체적으로 명시하는 방법에 대해 살펴보고자 한다.

현장 가격수집에서, 제품 가격은 일반적으로 특정 제품의 특정 시점, 특정 아울렛에서 관찰된다. 스캐너 데이터에서 제품 가격을 관찰하는 것이 아니라 품목코드의 단위 값(판매액을 판매수량으로 나눈 값)을 산출한다. 개별 제품을 지정할 때는 (1) 시간, (2) 대상처, (3)제품 3가지 차원을 고려해야 한다. 그림 1이 강조하는 바와 같이 각 수준에서 평균 가격(단위 값)이 계산된다. 지출액을 관련 수량으로 나누어 단가를 산정한다.

그림 개별제품 가격 도출(From product-offers to homogeneous products)[4)]



4) Index Compilation Techniques for Scanner Data, Claude Lamboray, UNECE ,Online meeting, June 2021

집계 순서(처음에는 시간, 나중에는 아울렛과 제품에 대한)는 중요하지 않으나, 단위값 편향의 가능성이 있으므로 이 순서로 3가지 차원을 고려한다.

개별 제품은 이러한 연속적 집계 수준에서 정의할 수 있고, 주어진 시간 동안 하나의 아울렛의 하나의 품목코드를 언급하면 매우 좁은 관점에서 정의될 수 있다. 더 넓은 개별 제품이라는 것은 시간이 지남에 따라 매칭을 증가시키는 것이다. 더 많은 데이터가 함께 묶일 때 지수 산출에서 고려될 개별 제품의 수는 감소할 것이다.

개별 제품들 규격은 데이터 특성(더 집계된 수준에서 제공)과 제품 범주(예: 특히 의류 관련)에 따라 크게 달라진다. 개념적으로, 주요 원칙은 거래/제품(product-offer)들은 이들 사이에 유의적 품질 차이가 없는 한 결합될 수 있다. '동종 제품'은 의미 있는 품질 차이가 없는 제품 집합으로 정의된다. 품질 차이는 이미 언급된 시간, 대상처 및 제품과 관련하여 평가된다. 품질이 다른 거래가 결합되면 단가 편향이 발생할 수 있다. 개별 제품의 사양에 대한 몇 가지 예는 표 1과 같다.

표1 개별 제품들의 규격(specification for the individual products) 예시[5)]

시간 범위	대상처 차원	제품 차원
해당월 첫 3주	소매체인별 단위 값	SKU
해당월 첫 14일	소매체인 및 지역별 단위 값	Article code
해당월 첫 2주	소매체인 및 상점유형별 단위 값	GTIN
해당월 첫 3주	소매체인 및 지역별 단위 값	GTIN
전체 해당월	대상처별 단위 값	GTIN
전체 해당월	대상처별 단위 값	동일품질로 구성된 GTIN그룹

시간 범위

원칙적으로 품목코드가 다른 소비자에게 다른 가격에 판매될 때, 같은 달 안에서 다른 시기에 판매할 때는 단가를 계산하는 것이 적절하다.
데이터 공급 체계와 생산 및 공표 일정에 따라, 개별 제품은 일반적으로 매월 2주, 3주(가끔 4주)범위이고, 기준월은 최대한 커버하는 것이 중요

5) Index Compilation Techniques for Scanner Data, Claude Lamboray, UNECE ,Online meeting, June 2021

하다. 아래 예제에서는 데이터가 주별로 제공된다. 4주는 한 달로 통합되고, 다른 품질 측면을 고려할 필요가 없다면, 단순히 전체 월의 평균 가격을 집계한다.

(시간차원 집계)[6]

품목코드	대상처	시간	매출액	판매수량	가격
1	1	1 주차	120	56	2.14
1	1	2 주차	130	63	2.06
1	1	3 주차	100	43	2.33
1	1	4 주차	200	120	1.67
집계					
1	1	1~4 주간	120+130+100+200=550	56+63+43+120=282	550/282=1.95

대상처 차원

아울렛 간 품질 차이는 다양한 개장시간 및 제품 구색 등과 관련될 수 있다. 같은 체인 아울렛이라도 가격전략이 다를 수 있고, 이상적으로는 개별 제품을 대상처 수준에서 데이터를 가능한 한 세분화하는 것이다. 실제로, 같은 소매 체인이나 브랜드 아울렛을 결합하는 데에는 이유가 있을 수 있다. 다른 아울렛에서 판매되는 동일한 제품이 아울렛 전체에 걸쳐 제품 가격이 체계적으로 같은 것으로 확인되면 동질적으로 볼 수 있다. 여러 아울렛을 종합하는 것은 지수 산출에 사용될 개별 상품 수를 감소시켜 계산 시간을 줄이고 데이터 저장 용량을 적게 요구한다. 대상처에 걸친 집계 여부의 영향은 평가할 수 있는 경험적 문제이다. 최종 지수에 대한 대상처 간 집계(또는 그렇지 않음) 영향을 테스트할 수 있다. 이것은 또한 데이터에서 이용할 수 있는 세부 정보의 수준에 따라 달라진다. 예를 들어, 데이터는 전체 소매점으로만 제공되어서 대상처별로 세분화할 수 없다.

(대상처별 집계)[6]

6) Guide on the use of multilateral methods in the HICP Draft version, October 2021

품목코드	대상처	시간	매출액	판매수량	가격
1	1	1~4 주간	550	282	1.95
1	2	1~4 주간	2203	1123	1.96
집계					
1	1 & 2	1~4 주간	550+2203 =2753	282+1123 =1405	2753/1405 =1.96

제품(product) 차원

GTIN은 데이터에서 가장 세분화된 제품 레벨인 경우로 알려져 있다. 이외에 일부 소매업체는 GTIN보다 약간 안정적일 수 있는 SKU도 사용할 수 있고 다른 제품 식별자가 있을 수 있다. 엄격한 제품 코드를 추적할 때 재출시 문제가 발생할 있다. 따라서 '동일한' 제품을 나타내는 재출시를 식별하고 두 품목 코드를 함께 연결하는 절차를 개발하는 것이 선호된다. 재출시는 평균 가격 계산 시 조정해야 하는 패키지 크기의 변화 일 수 있다. 품목 코드에서 재출시와 높은 회전율(제품 이탈)을 처리하기 위해 (i) 연결, (ii) 그룹화 또는 (iii) hedonic 전략을 사용할 수 있다.

(i) 연결(Linking)

대부분의 슈퍼마켓 제품의 경우에는, 데이터에서 사용할 수 있는 품목 코드(예: GTIN, SKU)를 제품 식별자로 사용하는 것으로 충분하고, 재출시를 식별하는 일부 절차와 이상적으로 결합되어야 한다. 이것은 예시[7]로 가장 잘 설명된다. 첫 번째 품목이 기간 1에서 20의 가격을 가지고 기간 2이후로는 사용할 수 없고, 동일한 품질의 두 번째 품목('재출시')의 가격은 25이며, 기간 2이후에만 구입할 수 있다. 매칭물가지수는 가격인상(20 → 25)을 놓치게 될 것이다. 이를 피하려면 두 제품이 연결돼야 한다.

	P1	P2	
GTIN A	20	-	비매칭으로 가격 변동이 추적되지 않는다.
GTIN B	-	25	
연결	20	25	가격 변동이 추적된다.

7) Guide on the use of multilateral methods in the HICP Draft version, October 2021

(ii) **그룹화(Grouping)**

균질한 제품을 구성할 때 균질성과 시간에 따른 안정성 사이에 절충이 이루어져야 한다. 동질 제품을 너무 광범위하게 정의하면 단위 값 편향이 발생할 수 있다. 너무 엄격하면 재출시가 캡처되지 않을 수 있다. MARS방법은 이 두 목표 사이의 타협점을 찾는 도구로 사용될 수 있다. 일반적으로 잦은 가격 인하로 끝나는 라이프 사이클 제품은 품질 차이가 작은 경우에 지속 기간이 긴 균질 제품으로 결합하는 것이 좋다(예 :특히, 의류)

(iii) **헤도닉**

제품 특성을 사용하여 균질제품을 형성하고 시간이 지남에 따라 매치를 하는 대신, 좁게 정의된 개별제품을 사용하고 헤도닉으로 조정된 다변지수 방법들에 제품특성을 직접 통합할 수 있다. 그룹화 할 것인지 아닌지는 제품 종류에 따라 결정되고 예를 들어, 의류의 경우 균질 제품을 만드는 것이 더 적합할 수 있지만, 전자 제품의 경우 헤도닉 추정이 더 일반적이다.

MARS 방법(Chessa, 2018) **사례**

MARS 방법은 동질 제품의 구성은 균질성과 제품 매칭이라는 두 가지 기준 사이의 절충이라는 생각에 기초한다. 이 두 가지 기준에 대한 메트릭이 제공되고, '최적' 제품 계층을 찾을 수 있다. 이 방법에는 제품을 설명하는 범주형 변수가 필요하다.

솔류션	품목코드	원단	소매	P0	Q0	P1	Q1
1	1	cotton	short	1.90	103	2.00	53
	2	organic	short			7.00	29
	3	cotton	long	14.00	15	15.11	18
	4	cotton	short	2.00	108	2.00	1
	5	cotton	short			5.10	50
2		cotton	short	1.95	211	3.49	104
		cotton	long	14.00	15	15.11	18
		organic	short			7.00	29
		organic	long				
3			short	1.95	211	4.26	133
			long	14.00	15	15.11	18
4		cotton		2.75	226	5.20	122
		organic				7.00	29
5				2.75	226	5.55	151

두 기간 티셔츠 데이터에 대한 위의 표 예를 생각해 보자. 데이터에는 5개 품목 코드가 있다. 원단(cotton/organic)과 소매(short/long)의 두 가지 제품 특성이 있다. 동종 제품을 정의하는 방법은 여러 가지가 있다. 각 솔루션에 대해 위 표와 같이 제품 매칭 및 균질성이 측정된다. 위 표와 같이 광범위한 균질 제품 정의를 가진 다섯 가지 다른 솔루션이 있다. 솔루션 1이 가장 세분화된 것이지만 솔루션 5에는 제품('티셔츠')이 하나만 있다. 기간 1의 티셔츠 수량은 5개의 품목 코드 수량의 합 : 53+29+18+1+50=151. 기간 1의 티셔츠 가격은 5개 품목 코드 평균 가격 : (2.00*53+7.00*29+15.11*18+2.00*1+5.10*50)/(53+29+18+1+50)=5.55

MARS를 사용하여 각 솔루션이 평가될 수 있다. 솔루션 3이 가장 높은 점수를 주는 것으로 나타나, 균질 제품이 소매별로 정의되어야 한다는 것을 제안한다.

실제 데이터에 대한 이 방법은 이 예시에서 제시된 것보다 더 복잡할 수 있다. 분석은 2시점 이상의 기간에 걸쳐 수행되어야 한다. 최적의 제품 정의는 선택한 기간에 따라 달라질 수 있다. 더욱이, 결과는 분석에 사용된 특정 데이터(제품 샘플 및 가격)에 지나치게 민감할 수 있다. MARS 방법은 의사 결정 보조 수단으로 간주되어야 한다.

솔루션 1: 제품이 '품목코드'로 정의
총 제곱합:
$53*(2.00-5.55)^2+29*(7.00-5.55)^2+18*(15.11-5.55)^2+1*(2.00-5.55)^2+50*(5.10-5.55)^2 = 2397.1$
관측된 제곱합:
$53*(2.00-5.55)^2 +29*(7.00-5.55)^2+18*(15.11-5.55)^2+1*(2.00-5.55)^2+50*(5.10-5.55)^2 = 2397.1$
균질성: 2397.1/2397.1 = 100% 상품매치 : (53+18+1)/151=47.7%
MARS 점수: 100%*47.7%=47.7%

솔루션 2: 제품이 '원단'과 '소매'로 정의, 총 제곱합: 2397.1 (솔류션 1 참조)
관측된 제곱합: $104*(3.49-5.55)^2+18*(15.11-5.55)^2+29*(7.00-5.55)^2=2147.6$
균질성: 2147.6/2397.1 = 89.6% 상품매치 : (104+18)/151=80.8%
MARS 점수: 89.6%*80.8%=72.4%

솔루션 3: 제품이 '소매'로 정의, 총 제곱합: 2397.1 (솔류션 1 참조)
관측된 제곱합: $133*(4.26-5.55)^2+18*(15.11-5.55)^2=1868.3$
균질성: 1868.3/2397.1 = 77.9% 상품매치: (133+18)/151= 100%
MARS 점수: 77.9%*100%=77.9%

솔루션 4: 제품이 '원단'으로 정의, 총 제곱합: 2397.1 (솔류션 1 참조)
관측된 제곱합: $122*(5.20-5.55)^2+29*(7.00-5.55)^2=75.5$
균질성: 75.5/2397.1 = 3.1% 상품매치 :122/151= 80.8%
MARS 점수: 3.1%*80.8%=2.5%

솔루션 5: 제품은 '티셔츠'로 정의, 총 제곱합: 2397.1 (solution 1 참조)
관측된 제곱합: $151*(5.55-5.55)^2=0$, 제품 균질성: 0/2397.1=0%
상품매치 : 151/151 = 100%, MARS 점수: 0%*100%=0%

5. CPI 산출 방법

5.1. 산출방법 프레임워크

개별제품 가격들을 집계하여 기본물가지수를 얻기 위한 여러 방법이 있고, 어떤 방법이든 개별제품의 가격 및 수량을 활용한다. 물가 지수를 구현하기 위한 여러 전략이 있고, (i) 가중치와 제품군, (ii) 지수 특성, (iii) 업데이트 및 연결, (iv) 샘플링 방법에 따라 산출 방법을 구분한다.
이 프레임워크는 그림1에 설명되어 있다.

가중치 및 제품군(Weights and product universe)

각 기간에서 사용 가능한 개별 제품(individual product) 세트가 다를 수 있다. 이 세트는 개별 제품 정의에 따라 달라지고, 이로 인해 다른 제품군이 발생할 수 있다. 정적 제품군, 동적 제품군 그리고 가중치 관련 개별 제품들 간의 주요 구별로 접근이 이루어질 수 있다.

지수 속성(Index property)

물가지수가 충족하거나 충족하지 못하는 속성[8]들이 있다. 다른 산출 방법과 구별하기 위해 이행성, 시점역전성 및 동일성에 주요 초점을 맞춘다.

업데이트 및 연결 전략(Update and linking strategy)

양변지수의 경우, 기준 기간은 개별 제품들의 기본 바스켓과 함께 업데이트된다. 이 문제는 매달 발생할 수 있으며, 예를 들어 1년에 한 번만 발생할 수 있다. 자주 업데이트하면 동적 제품 환경이 고려되고, 체인을 자주하면 연쇄 편의가 발생할 수 있다. 다변지수를 사용하면 지수산출기간(time window)이 매월 업데이트되고, 대부분의 경우 한 달간 앞당겨지는 롤링 타임 윈도우가 사용된다. 짧은 실행 기간(time windows) 동안 산출된 지수의 연결은 불안정한 결과를 초래할 수 있으며 연쇄 편의 문제를 해결

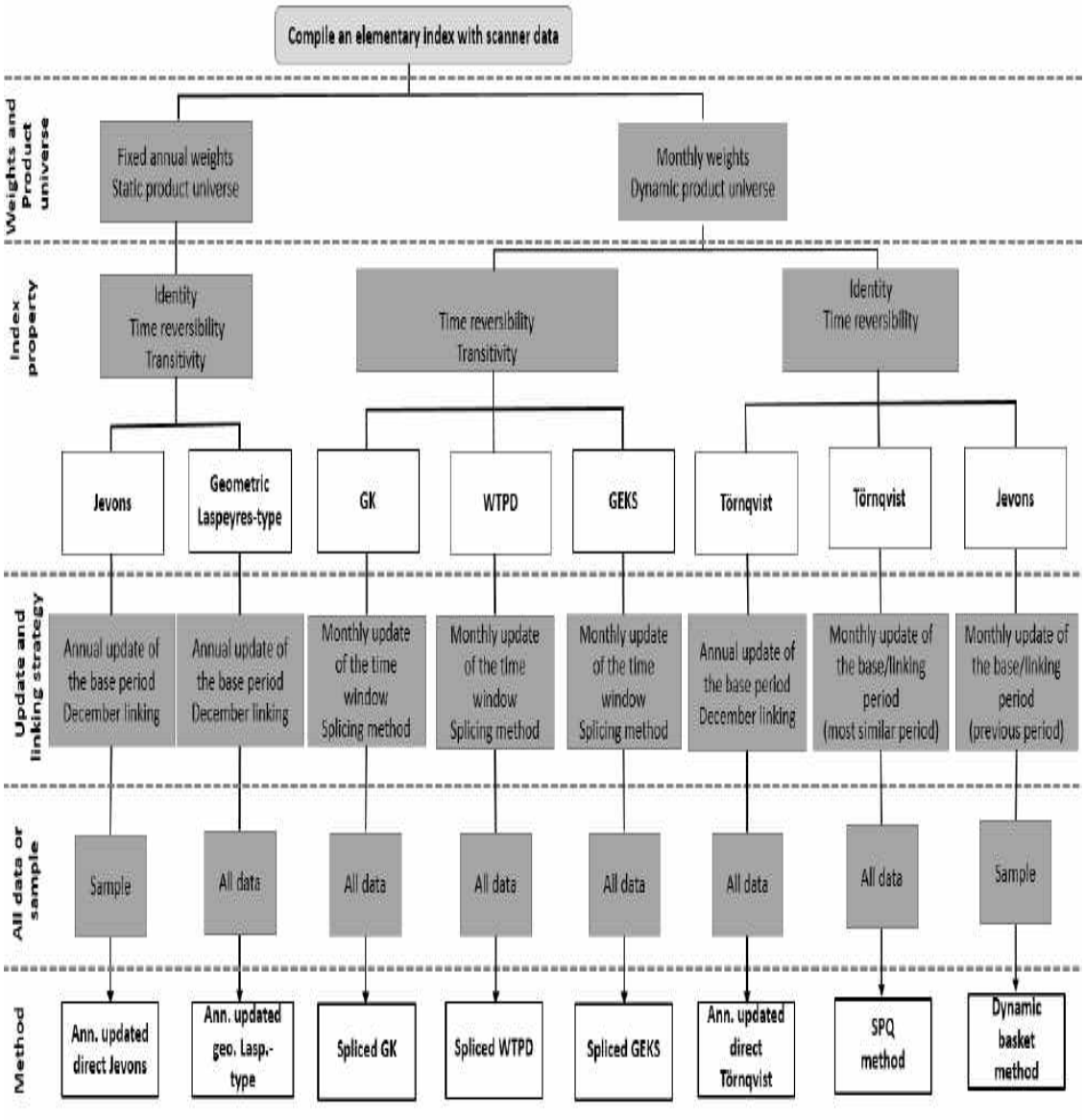
8) 동일성, 이행성은 다음 지수시험 접근 설명을 참고하고, 시점역전(time reversibility)은 (a)와 (b) 사이의 지수가 (b)와 (a) 사이의 동일한 지수의 역과 같아야 하는 속성 $P(p^1, q^1, p^2, q^2) = \frac{1}{P}(p^2, q^2, p^1, q^1)$ 이다.

하지 못할 수 있고, 실행 기간이 길어지면 과거의 더 많은 데이터 수가 현재 월 산출에 영향을 미친다.

모든 데이터 또는 표본 추출(all data or sampling)

제품군의 모든 데이터가 지수 산출에 이용되거나, 제품군을 대표하고 지수에 들어가는 개별제품을 제한하는 표본이 선택될 수 있다.

<그림1> 지수 산출방식 프레임워크[9)]



9) Index Compilation Techniques for Scanner Data, Claude Lamboray, UNECE ,Online meeting, June 2021

5.2. 물가지수 산식

현재 국가들이 활용하거나 검토 중인 양변지수와 체인, 다변지수를 살펴보고자 한다.

스캐너 데이터는 기존 샘플 기반 방법을 사용하여 수행될 수 있다. 이전에 가격 수집가들이 대상처를 방문하는 가격은 샘플링 설계나 지수산식을 변경하지 않고 스캐너 데이터의 단위 값으로 대체될 수 있다. 표본 추출하는 대신 사용가능한 모든 데이터를 사용하기로 결정한 경우, 다변지수가 선호 될 수 있다. 이는 원래 여러 국가의 가격 수준을 비교하기 위해 개발되었지만, 시간 경과에 따른 가격 비교에 적용할 수 있고, 상품회전율[10]이 높고 판촉 판매가 자주 발생하는 스캐너 데이터에 특히 유용하다. ILO 및 Eurostat 매뉴얼에서는 스캐너 데이터 산출방법에 있어 다변지수 방법을 가장 중요시 다루고 있다.

5.2.1. 양변지수

먼저, 양변지수에 대해 정적 및 동적 제품군으로 나누어 살펴보고, 그 다음에 다변지수에 대해 좀 더 자세히 파악해 보고자 한다. 양변지수 정적 제품군에서는 고정가중치 제본스와 기하 라스파이레스가 있다.
제본스 지수를 활용하는 국가는 덴마크, 스웨덴, 아이슬란드, 스위스 등이 있고, 기하 라스파이레스 지수는 프랑스 등에서 사용하고 있다.

Jevons $$I_J^{0:t} = \prod_{i \in S} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{|S|}} = \frac{\prod_{i \in S} \left(p_i^t \right)^{\frac{1}{|S|}}}{\prod_{i \in S} \left(p_i^0 \right)^{\frac{1}{|S|}}}$$

모든 개별 제품 N에 대한 기하 라스파이레스 지수를 계산하는 대신, 제본스는 작은 표본 S에 대해 상대 가격의 기하 평균으로 계산할 수 있다. 표본 S는 기준 기간에 대해 선택되고 시간에 따라 가격이 측정된다. 시간 경과에

10) turnover : the rate at which goods are sold and replaced in a store

따라 샘플 S는 수동으로 유지할 수 있도록 상대적으로 작다. 지출 정보가 없거나 부족한 상황에서는 제본스 산식 사용을 권장될 수 있다.

기하 라스파이레스

$$I_{GL}^{0:t} = \prod_{i \in N} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{p_i^b q_i^b}{\sum_{j \in N} p_j^b q_j^b}}$$

신상품들은 사라지는 상품들의 일대일 대체로 지수 산출에 지속적으로 통합될 수 있다. 이용 가능한 제품이 없어지면 교체제품을 선택해서 지수에 가져와 품질조정을 수행한다. 이렇게 하면 처음에 선택한 바스켓이 고정된 상태로 유지되고 대표 상태를 유지할 수 있다. 다만 소멸 상품 대체로 사용되지 않는 개별 신상품은 무시된다.

양변지수 동적 제품군에서는 고정기반 Törnqvist와 동적바스켓 지수가 있다. 고정기반 Törnqvist지수를 활용하는 국가는 핀란드, 동적바스켓은 이태리, 스페인, 슬로베니아 등에서 사용하고 있다.

고정기반 Törnqvist

스캐너 데이터는 품목에 대한 지출정보를 포함하기에 최상급 가격 지수 구성이 가능하다. Fisher와 Törnqvist는 유사한 결과를 도출하지만, Törnqvist는 더 간단한 표현을 가능하게 하고, 실제로 사용된다. Törnqvist 지수는 다음과 같이 정의된다.

$$I_T^{0:t} = \prod_{i \in S} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}}, \; s_i^0 = \frac{p_i^0 q_i^0}{\sum_{i \in S} p_i^0 q_i^0} \text{ and } s_i^t = \frac{p_i^t q_i^t}{\sum_{i \in S} p_i^t q_i^t}$$

이 지수는 동일성과 시간가역성을 만족하지만 이행적이지 않다. 이 지수 월간연쇄는 '연쇄편의' 문제를 초래할 수 있으므로 피하는 것이 최선이다. 현재기간을 고정 기준기간과 직접 비교시 체인이 관련되지 않기에 이 문제를 피할 수 있다. 또한 개별제품에 부여된 변동 가중치는 각 기간별

개별제품의 경제적 중요성을 파악할 수 있다. 두 비교기간의 상품 중복을 극대화하기 위해 1년 전체(계절상품을 통합)를 기준기간으로 사용할 수 있다. 실제로 13개월(y-1년 12월에서 y년 12월) 고정기준 물가지수가 작성된다. 각 달은 y-1년과 비교된다. 다음 해에는 기준기간이 업데이트되어 y년이 새로운 기준기간이 되고 13개월(y년 12월부터 y+1년 12월까지) 동안 고정 기준지수가 작성된다. 두 시리즈는 y년의 12월을 중첩기간으로 사용하여 연결된다.

표1은 연쇄 Törnqvist 하향 움직임 예를 보여준다. 2개 품목과 9개 시점이 있다. 품목 1과 2의 '정규'가격은 각각 3과 4이지만, 품목 1은 시점 3과 7이, 품목 2는 시점 2와 6이 한시적으로 인하된다. 마지막 시점 9에는 가격과 수량이 첫 번째 기간의 것과 동일하다. 그럼에도, 기간별 연쇄 Törnqvist 물가지수는 78.18에 그쳐, 거의 22%의 가격이 하락하였다. 다변지수와 직접 Törnqvist의 경우 지수는 첫 번째 기간과 같다.

표1 연쇄편의 예시[11)]

시점	p1	p2	q1	q2	비중1 (%)	비중2 (%)	품목1 t와 t-1 평균비중	품목2 t와 t-1 평균비중	Tornqvist 단기지수	Tornqvist 연쇄지수
1	3	4	12	10	47.4	52.6				100
2	3	2	12	30	37.5	62.5	42.4	57.6	67.10	67.10
3	1	4	40	5	66.7	33.3	52.1	47.9	78.66	52.78
4	3	4	5	10	27.3	72.7	47.0	53.0	167.53	88.42
5	3	4	12	10	47.4	52.6	37.3	62.7	100.00	88.42
6	3	2	12	30	37.5	62.5	42.4	57.6	67.10	59.33
7	1	4	40	5	66.7	33.3	52.1	47.9	78.66	46.67
8	3	4	5	10	27.3	72.7	47.0	53.0	167.53	78.18
9	3	4	12	10	47.4	52.6	37.3	62.7	100.00	78.18

동적 바스켓 방법(dynamic basket)

할인판매로 인한 연쇄편의를 최소화 하는 방법은 품목 가중치를 매기지 않고 기간별 매칭인 Jevons를 연쇄하는 것이다. 여기서 $N_M^{t-1,t}$은 기간 t-1과 t 사이의 매칭 품목수이다.

11) International Labour Office. (2020). "Consumer Price Index Manual: Theory and Practice." Geneva. p228

월별 제본스(단기지수) $I_J^{t-1:t} = \prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{p_i^{t-1}} \right)^{\frac{1}{N_M^{t-1,t}}}$

슈퍼마켓 스캐너 데이터 처리(Eurostat, 2017)에서 '동적 바스켓 방법'은 매월 기준기간을 갱신하는 것으로 매월 지수편성에 들어가는 개별 상품 세트가 재샘플링된다. 실제로 두 기간 연속 가장 많이 팔린 제품을 선정하는 컷오프 샘플링이 적용된다. 이 지수는 선택된 개별상품만을 고려하여 두 기간에 작성되고, 최종지수는 월별 제본스를 연쇄하여 구한다. 두 달 연속 계산되는 지수는 동일성과 시점 역전성을 만족하나, 다른 바구니에 걸쳐 계산된 연쇄 제본스는 더 이상 이행성이지 않다. 연쇄 매칭 제본스 지수활용이 문제가 없다는 것은 아니다. 예를 들어, 염가처분 판매는 지수에 하방 압력을 가할 수 있고, 이 문제를 완화하기 위해 가격과 판매 수량이 모두 급격히 떨어진 품목을 제거(예: 의류)하는 덤프필터를 사용할 수 있다.

가중치 결여도 문제가 있다. 제품지출은 대개 매우 치우쳐 있으므로, 많은 저지출 제품은 소수 고지출 제품과 동일한 가중치를 부여받게 될 것이다. 대략적인 암묵적 가중치는 임계치를 사용하여 저지출 제품을 제외함으로써 얻을 수 있다. 이 방법은 가중치가 제품샘플링에 사용되지만 지수계산에서는 명시적으로 사용되지 않기 때문에 연쇄편의의 위험을 감소시킨다. 이는 데이터에 포함된 정보를 최적으로 활용하면서 동시에 일반적인 양변지수라는 장점이 있다. 따라서 이 방법은 사용자에게 설명하기가 쉽다. 그러나 이는 최적은 아니고, 보다 발전된 해결책은 제품에 명시적으로 가중치를 부여하고 다변지수를 구성하는 것이다.

5.2.2. 다변지수[12)]

이 방법은 변동가중치 및 동적 제품군에 해당되며 이행성을 충족한다. 체인으로 작성될 수 있으며, 연쇄편의가 없다. 주로 다음과 같은 세 가지 유형으로 설명할 수 있다. 현재 유로지역에서 네덜란드. 벨기에, 룩셈

12) Guide on the use of multilateral methods in the HICP Draft version(October 2021), ILO(2020). "Consumer Price Index Manual: Theory and Practice." 10. Scanner data. 사례를 가지고 설명

부르크, 노르웨이가 이 방법을 활용하여 지수를 생산하고 있다.

- GEKS 지수는 time window에 속하는 두 기간을 비교하는 데 사용되는 양변지수를 기반으로 지수를 평균한다. 주로 Törnqvist와 함께 적용된다.
- Geary-Khamis(GK) 지수는 금액(value)지수를 수량지수로 나눈 값 지수로 정의되는 암묵적 물가 지수로 볼 수 있다.
- WTPD 지수는 time window에 속하는 제품 및 시간에 대한 더미 변수를 활용하여 회귀분석을 하는 것으로, 각 관측치는 주어진 기간 t의 비중에 따라 가중치가 부여되는 가중 최소 제곱(WLS)을 사용한다.

세 지수는 모두 이행성(또한 시간역전성 충족) 충족하지만 동일성 검정을 만족시키지 못한다. 다변지수가 적용되는 time window 길이에 대해 결정되어야 하고, 새로운 time window이 사용될 때마다 이전지수가 변경될 수 있다. 따라서 이미 발표된 지수 수정을 피하기 위해 최신 다변지수를 이전 결과에 연결하는 스플라이싱 기법을 사용해야 한다. 결국 발표된 지수에 대해서 이행성이 더 이상 충족되지 않는다. 따라서 스플라이싱된 지수에 대해 일정 정도의 연쇄편의를 완전히 제외할 수 없다. 가중치 없이 구현될 수도 있으나, 가격과 수량 모두에 의존한다. 각 다변지수는 4개 제품과 시점으로 구성된 다음 예시로 설명할 수 있다.

개별제품	p0	q0	p1	q1	p2	q2	p3	q3
1	2.97	15	2.96	25	2.93	32	3.03	33
2	3.64	44	3.50	79	3.36	65	3.42	90
3	6.75	49	6.71	41	6.67	35	6.73	53
4	3.37	35	3.29	59	3.37	30	3.37	31

GEKS(Gini-Eltetö-Köves-Szulc)

GEKS는 time window에 속하는 두 기간을 비교하는 데 양변지수를 기반으로 한다. 0과 t사이 지수가 주어진 time window W에 대해 양변지수가 시간역전성을 만족한다면, GEKS 지수를 다음과 같이 쓸 수 있다.

$$I^{0:t}_{W(GEKS-Tq)} = \prod_{k\in W}\left(\frac{I^{0,k}_{Tq}}{I^{t,k}_{Tq}}\right)^{\frac{1}{|W|}} \equiv \prod_{k\in W}\left(I^{0,k}_{Tq} * I^{k,t}_{Tq}\right)^{\frac{1}{|W|}}$$

첫 번째 단계는 Törnqvist를 계산하는 것이다. 예제 데이터 세트에 대한 이러한 2x2 비교 계산 결과는 다음과 같은 행렬로 요약할 수 있다.

시점	0	1	2	3
0	1.0000	0.9810	0.9705	0.9820
1	1.0194	1.0000	0.9877	0.9998
2	1.0304	1.0124	1.0000	1.0140
3	1.0184	1.0002	0.9862	1.0000

예를 들어, 시점 0과 비교한 시점 1의 Törnqvist는 0.9810이고. 이 행렬 한 쪽 대각선 값은 1이고 행렬은 대칭이다. 시점 1과 비교한 시점 0지수는 시점 0과 비교한 시점 1의 지수의 역수와 같다(1.0194=1/0.9810). GEKS-Tq 지수는 다음과 같이 도출할 수 있다.

$$I^{0,1}_{[0,3]} = ((I^{0,0}_{Tq} * I^{0,1}_{Tq})(I^{0,1}_{Tq} * I^{1,1}_{Tq})(I^{0,2}_{Tq} * I^{2,1}_{Tq})(I^{0,3}_{Tq} * I^{3,1}_{Tq}))^{1/4}$$

= ((1*0.9810)(0.9810*1)(0.9705*1.0124)(0.9820*1.0002))$^{1/4}$ = 0.9817

$$I^{0,2}_{[0,3]} = ((I^{0,0}_{Tq} * I^{0,2}_{Tq})(I^{0,1}_{Tq} * I^{1,2}_{Tq})(I^{0,2}_{Tq} * I^{2,2}_{Tq})(I^{0,3}_{Tq} * I^{3,2}_{Tq}))^{1/4}$$

= ((1*0.9705)(0.9810*0.9877)(0.9705*1)(0.9820*0.9862))$^{1/4}$ = 0.9696

$$I^{0,3}_{[0,3]} = ((I^{0,0}_{Tq} * I^{0,3}_{Tq})(I^{0,1}_{Tq} * I^{1,3}_{Tq})(I^{0,2}_{Tq} * I^{2,3}_{Tq})(I^{0,3}_{Tq} * I^{3,3}_{Tq}))^{1/4}$$

= ((1*0.9820)(0.9810*0.9998)(0.9705*1.0140)(0.9820*1))$^{1/4}$ = 0.9822

기어리-카미스((Geary-Khamis, GK)

GK는 품질조정 값 지수[13]의 한 가지 예이고, 다음 식을 풀어서 얻어진다. 아래 (1)식 분자는 표본기간에 걸쳐 고정된 "기준가격" v_i를 가진 물가지수(기간 t 수량 사용)이다. **지수는 시작 기간 0에서 1이다.**

$$I^{0,t}_{W(GK)} = \frac{\sum_{i\in S^t} p_i^t q_i^t / \sum_{i\in S^0} p_i^0 q_i^0}{\sum_{i\in S^t} v_i q_i^t / \sum_{i\in S^0} v_i q_i^0} = \frac{[\sum_{i\in S^t} S_i^t [p_i^t / v_i]^{-1}]^{-1}}{[\sum_{i\in S^0} S_i^0 [p_i^0 / v_i]^{-1}]^{-1}} \quad (1) \quad v_i = \sum_{z\in W} \frac{q_i^z}{\sum_{s\in W} q_i^s} \frac{p_i^z}{I^{0,z}_{W(GK)}} \quad (2)$$

13) 네덜란드 CPI에서 이 방법의 실제 적용에 대해서는 Chessa(2016)를 참조.

기준가격(v_i)은 위 식 (2)와 같고, 전체 표본기간에 걸쳐 품목의 총 판매수량의 각 기간 비중이 가중치 역할을 하는 디플레이트된 관측 가격의 가중 산술평균과 같다. GK는 식 (2)에서 디플레이터 역할을 하기에, 식 (1)과 (2)는 동시에 풀어야 하는 식 체계이다. 다음은 예제 데이터에 대한 GK이다. 4개 제품에는 v1= 3.029974, v2= 3.523584, v3=6.820025, v4=3.392614의 조정 계수(기준가격)[14]가 사용된다.

시점	금액(value)	볼륨(Volume)	가격 지수
0	2.97*15+3.64*44+6.75*49+3.37*35=653.41	3.029974*15+3.523584*44+6.820025*49+3.392614*35=653.41000	(653.41/653.41)/(653.41/653.41)=1.0000
1	2.96*25+3.50*79+6.71*41+3.29*59=819.72	3.029974*25+3.523584*79+6.820025*41+3.392614*59=833.8977124	(819.72/833.8977124)/(653.41/653.41)=0.9830
2	2.93*32+3.36*65+6.67*35+3.37*30=646.71	3.029974*32+3.523584*65+6.820025*35+3.392614*30=666.4713966	(646.71/666.4713966)/(653.41/653.41)=0.9703
3	3.03*33+3.42*90+6.73*53+3.37*31=868.95	3.029974*33+3.523584*90+6.820025*53+3.392614*31=883.7440252	(868.95/883.7440252)/(653.41/653.41)=0.9833

각 개별 제품의 시점별 수량기반 비중은 다음과 같이 계산할 수 있다.

개별제품	전체기간	시점1	시점2	시점3	시점4
0	15+25+32+33=105	15/105=14.3%	25/105=23.8%	32/105=30.5%	33/105=31.4%
1	44+79+65+90=278	44/278=15.8%	79/278=28.4%	65/278=23.4%	90/278=32.4%
2	49+41+35+53=178	49/178=27.5%	41/178=23.0%	35/178=19.7%	53/178=29.8%
3	35+59+30+31=155	35/155=22.6%	59/155=38.1%	30/155=19.4%	31/155=20.0%

이제 조정 계수(기준가격)는 이전 단계에서 얻은 가격 지수에서 도출될 수 있다. 각 기간의 가격은 물가지수에 의해 디플레이트 된다. 4개 제품의 전체 기간 동안의 평균 가격은 수량 비중을 사용하여 얻어진다.

v1=14.3%*2.97/1+23.8%*2.96/0.9830+30.5%*2.93/0.9703+31.4%*3.03/0.98273=**3.029974**
v2=15.8%*3.64/1+28.4%*3.50/0.9830+23.4%*3.36/0.9703+32.4%*3.42/0.98273=**3.523584**
v3=27.5%*6.75/1+23.0%*6.71/0.9830+19.7%*6.67/0.9703+29.8%*6.73/0.98273=**6.820025**
v4=22.6%*3.37/1+38.1%*3.29/0.9830+19.4%*3.37/0.9703+20.0%*3.37/0.98273=**3.392614**

14) 실제로, 모든 i에 대해 vi = 1로 시작한 다음, solution으로 수렴하기 전에 추가로 반복할 수 있다.

기준가격은 처음에 사용된 것과 동일하고, 식 체계를 반복함으로서 풀 수 있다. 주어진 조정계수 집합으로 시작한다(예: 모든 제품에 대해 vi=1). 조정계수와 지수 값이 더 이상 변경되지 않으면 반복이 중지된다.

3) 시간 제품 더미(weighted time product dummy method, WTPD or TPD)

$$\ln p_i^t = \alpha + \sum_{t=1} \delta^t D_i^t + \sum_{i=2} \gamma_i K_i + \varepsilon_i^t$$

여기서 K_i는 개별제품 i와 관련이 있는 경우 1이고 그렇지 않은 경우 0인 더미변수이고, D_i^t는 개별제품 i가 시점 t와 관련이 있는 경우 1이고 그렇지 않은 경우 0인 더미변수이다. 첫 번째 개별 제품과 첫 번째 시점 0의 더미는 모형을 식별하기 위해 제외된다. 회귀분석은 가중 최소 제곱(WLS) 사용하여 추정되며, 이 추정치는 잔차 제곱의 가중 합을 최소화한다. 각 관측치는 지정된 기간 t에서의 비중에 따라 가중치가 부여된다.

$$s_i^t = \frac{p_i^t q_i^t}{\sum_{i \in S} p_i^t q_i^t}$$

최종 지수는 시간 더미변수 추정치의 지수를 취하여 구한다.

$$I_{W(WTPD)}^{0,t} = \exp(\hat{\delta^t})$$

다음 표는 회귀 분석에 사용된 데이터이다. 회귀분석의 경우 가격그가 종속변수로 사용된다. 3개 제품 더미변수(K2, K3 및 K4)와 3개 기간 더미변수(D1, D2, D3)가 있다. 각 관측치는 해당 기간의 각 제품 지출 비율에 따라 가중치가 부여된다.

회귀 분석에서 얻은 추정치는 α=1.10754; δ1=-0.01743, δ2=-0.03019, δ3 =-0.01703, γ2=0.15287, γ 3=0.81206, γ4=0.11488이다. 지수는 시간 더미에 대한 추정치의 지수를 취하여 도출할 수 있다.

$$I_{[0,3](WTPD)}^{0,1} = \exp(\delta^1) = \exp(-0.01743) = 0.9827$$
$$I_{[0,3](WTPD)}^{0,2} = \exp(\delta^2) = \exp(-0.03019) = 0.9703$$
$$I_{[0,3](WTPD)}^{0,3} = \exp(\delta^3) = \exp(-0.01703) = 0.9831$$

ln(p)	D^1	D^2	D^3	K^2	K^3	K^4	가중치
1.088562	0	0	0	0	0	0	6.8%
1.291984	0	0	0	1	0	0	24.5%
1.909543	0	0	0	0	1	0	50.6%
1.214913	0	0	0	0	0	1	18.1%
1.085189	1	0	0	0	0	0	9.0%
1.252763	1	0	0	1	0	0	33.7%
1.903599	1	0	0	0	1	0	33.6%
1.190888	1	0	0	0	0	1	23.7%
1.075002	0	1	0	0	0	0	14.5%
1.211941	0	1	0	1	0	0	33.8%
1.89762	0	1	0	0	1	0	36.1%
1.214913	0	1	0	0	0	1	15.6%
1.108563	0	0	1	0	0	0	11.5%
1.229641	0	0	1	1	0	0	35.4%
1.905088	0	0	1	0	1	0	41.0%
1.214913	0	0	1	0	0	1	12.0%

5.2.3. 매칭 부족 및 품질 조정

ILO매뉴얼의 스캐너 데이터 명시적 품질조정에 관해 살펴보고자 한다. 명시적 품질조정을 허용하는 품목특성에 대한 데이터가 선호(헤도닉) 될 수 있는데, 유용한 시작은 다변 시간더미 헤도닉(TDH) 모델이다.

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t$$

위 z_{ik}은 품목(item) i에 대한 특성치 k(k =1,...K)이다. 앞서 설명한 TPD모델은 위 식 TDH로부터 도출되며, 품목지정 고정효과 $\exp(\gamma_i)$로 헤도닉 효과 $\exp(\sum_{k=1}^{K} \beta_k z_{ik})$를 대체한 것이다. TPD추정과 유사하게, 위 식은 전체 기간(t = 0,....,T) 데이터를 기초로 지출비중 가중회귀로서 추정된다.
다른 바코드나 SKU를 갖는 동질 상품 재출시가 낮은 매칭의 주요 원인이라면 바코드나 SKU가 아닌 그 특성에 따라 상품을 정의하는 것이 옵션이 될 수 있다. 그러나 스캐너 데이터에는 일반적으로 상당히 광범위한 품목 설명이 포함되어 있다. 이 설명에서 크기 및 브랜드 이름과 같은 몇 가지 특성만 추출할 수 있다. 이 경우 다양한 '그룹들'에 속하는 SKU 등 바코드에 걸쳐 단위값으로 계산되는 가격은 단위 가치 편향에 시달릴 수 있다.

GEKS가 TPD나 GK에 비해 유리한 점은 매칭되지 않은 새롭고 사라지는 품목의 “누락된 가격“이 대체(impute)될 수 있다는 것이다. 예를 들어, 헤도닉 대체 Törnqvist로 명시적 품질조정 GEKS를 추정할 수 있다. 이는 특성에 따라 품목(item)을 정의할 필요가 없다. 그러나, 중요한 특성에 대한 정보가 누락되면 누락 변수 편향이 발생할 수 있어 헤도닉 품질조정이 문제가 될 수 있다. 또한 "group 접근법"에서는 단위가치 편향을 야기할 수 있다. 일부 통계청에서는 스캐너 데이터를 풍부하게 하기 위해 소매업체 또는 제조업체 웹 사이트의 품질 특성을 관찰하기 위해 웹 스크래핑의 사용을 탐구해왔다. 시장조사회사에서 얻은 스캐너 데이터는 이미 품목특성에 대한 자세한 정보를 포함하고 있을 수 있다.

5.2.4. 다변지수 수정[15)]

다변지수는 새로운 데이터를 이용하면, 이전에 산출된 지수가 수정된다. 이는 CPI를 수정할 수 없기에 문제가 되고, 시간이 지남에 따라, 최근 물가 변동이 과거 가격 변화에 의해 영향을 받는다(특성 loss). 지수를 수정하지 않고 시계열를 연장하는 방법이 제안되었고, 윈도우조정(롤링 타임 또는 연장 윈도우), 스플라이싱 기법으로 특징 지어질 수 있다. 롤링 타임 윈도우는 매달, 시간 창이 한 달씩 앞당기고, 시간 창 길이는 일정하게 유지된다. 예를 들어, 13개월 롤링 타임 윈도우를 가정하면, 첫 번째 시간 창은 2019년 1월부터 2020년 1월이고, 그 다음은 2019년 2월부터 2020년 2월이다. 이 방법은 윈도우 길이를 일정하게 유지하면서 가장 오래된 달은 제거되고 최근 달은 포함된다. 시간 창 연장은 매달, 시간 창은 한 달씩 연장된다. 1년 후, 시간 창의 길이를 초기 길이로 재설정할 수 있다. 예를 들어, 첫 번째 시간 창은 2019년 12월부터 2020년 1월, 다음 시간 창은 2019년 12월부터 2020년 2월이다. 이 방법은 창 길이에 불균형이 생기고 연초 길이가 매우 짧다. 장점은 긴 과거 데이터 없이도 구현할 수 있다. 지수수정을 피하기 위해 지수를 이전 결과와 연결하는 스플라이싱 기법이 사용될 수 있다. 각 방법에 대해서 표1~5에서 설명하고자 한다.

- 이동 스플라이스(movement splice): t-1이 연결 시점

15) 표1~표5 출처 : Guide on the use of multilateral methods in the HICP Draft version (October 2021), ILO(2020). “Consumer Price Index Manual: Theory and Practice.” 10. Scanner data.

- 윈도우 스플라이스(window splice): t-T(윈도우 길이)+1이 연결 시점
- 하프 스플라이스(half splice) : t-(T+1)/2+1이 연결 시점
- 평균 스플라이스(mean splice): 모든 오버랩 기간이 연결 시점
- 고정 기준(fixed base): 이전 12월이 연결 시점

표1~4에서는 13개 시점 롤링 타임 윈도우가 사용된다. 13시점에 산출된 지수는 1-13시점을, 14시점 다변지수는 2-14 시점을 적용한다. 그런 다음 최신 산출결과가 이전에 계산된 지수에 연결된다. 예를 들어, 이동 스플라이스는 가장 최근 시점 변경 내용을 이전 기간에 링크한다. 표1에서 14시점(t) 지수는 시점 13과 14사이 새로 산출된 가격 변화를 이전에 계산된 지수(t-1)에 연결하여 구한다. 윈도우 스플라이스는 새로운 지수 변화를 이전에 계산된 지수 t-T(윈도우 길이)+1, 하프 스플라이스는 이전에 계산된 t-(T+1)/2+1이 연결시점으로 사용된다. 평균 스플라이스는 모든 링크(오버랩) 기간을 적용하여 얻은 가격 지수의 기하평균을 사용한다.(호주 ABS 적용방법). 마지막으로 고정기준은 전년 12월이 가격 기준기간으로 작용한다.

연속적인 윈도우 산출은 같은 기간 처음 발표된 지수의 재계산된 지수를 생성한다. 다시 계산된 그리고 공표된 지수는 모두 새로운 지수 계열이 연결될 수 있는 후보다. 이미 공표된 지수에 연결은 최근 chessa(2019년)에 의해 제안되었고, 장점이 있다. 예를 들어, 연결 시점이 12개월 전 시점에 해당하는 경우 새로운 창에서 계산한 전년대비 비율이 공표된 수치가 될 것이다.

표1 13개월의 롤링 윈도우, 이동 스플라이스 : t-1이 연결 시점

시점		1	2	3	4	5	…	11	12	13	14	15
1	13	100.0	100.7	100.6	101.6	102.7	…	104.3	106.0	103.8		
2	14		100.0	100.2	101.1	102.2	…	103.8	105.5	103.3	104.6	
3	15			100.0	101.0	102.0	…	103.5	105.3	103.2	104.4	104.1
공표지수		**100.0**	**100.7**	**100.6**	**101.6**	**102.7**	…	**104.3**	**106.0**	**103.8**	**105.1**	**104.8**

* 스플라이싱은 14시점에서 시작. 1차 산출에서 1-13시점에 대한 지수를 구한다. 시점 14 공표 지수는 2차 산출 시점 13과 14의 변동을 시점 13(103.8×104.6/103.3 = 105.1)의 지수에 적용하여 구한다. 시점 15 공표지수는 3차 산출 시점 14와 15 변동을 기간 14의 공표 지수(105.1×104 1/104.4 = 104.8)에 적용하여 구한다.

표2 13개월 롤링 윈도우, window splice : t-T(window 길이)+1이 연결 시점

시점		1	2	3	4	5	…	11	12	13	14	15
1	13	100.0	100.7	100.6	101.6	102.7	…	104.3	106.0	103.8		
2	14		100.0	100.2	101.1	102.2	…	103.8	105.5	103.3	104.6	
3	15			100.0	101.0	102.0	…	103.5	105.3	103.2	104.4	104.1
공표지수(이전시점 연결)		**100.0**	**100.7**	**100.6**	**101.6**	**102.7**	…	**104.3**	**106.0**	**103.8**	**105.3**	**105.0**
공표지수(기표지수 연결)		**100.0**	**100.7**	**100.6**	**101.6**	**102.7**	…	**104.3**	**106.0**	**103.8**	**105.3**	**104.7**

* **(이전시점에 연결)** 스플라이싱은 시점 14에서 시작. 1-13 시점 공표지수는 1차 산출에 의해 구한다. 시점 14 공표지수는 1차 산출 시점 13과 2 변동과 2차 산출 시점 2와 14 변동을 공표지수 시점 13지수에 적용하여 구한다(103.8×(100.7/103.8)×(104.6/100.0)=105.3). 시점 15 공표지수는 2차 산출 시점 14와 3의 변동과 3차 산출 시점 3과 15의 변동을 공표지수 시점 14 지수에 적용하여 구한다(105.3×(100.2/104.6)×(104.1/100.0)=105.0).

* **(공표지수에 연결)** 스플라이싱은 시점 14에서 시작. 1-13시점 공표지수는 1차 산출에 의해 구한다. 시점 14 공표지수는 2차 산출 시점 2와 14 변동을 공표지수 시점 2 지수에 적용하여 구한다(100.7 × 104.6/100.0 = 105.3). 시점 15 공표지수는 3차 산출 시점 3과 15 변동을 공표지수 시점 3 지수에 적용하여 구한다(100.6 × 104.1/100.0 = 104.7).

표3 13개월 롤링 윈도우, mean splice : 모든 오버랩 기간이 연결 시점

시점		1	2	3	4	5	…	11	12	13	14	15
1	13	100.0	100.7	100.6	101.6	102.7	…	104.3	106.0	103.8		
2	14		100.0	100.2	101.1	102.2	…	103.8	105.5	103.3	104.6	
3	15			100.0	101.0	102.0	…	103.5	105.3	103.2	104.4	104.1
공표지수(이전시점 연결)		**100.0**	**100.7**	**100.6**	**101.6**	**102.7**	…	**104.3**	**106.0**	**103.8**	**105.1**	**104.8**
공표지수(기표지수 연결)		**100.0**	**100.7**	**100.6**	**101.6**	**102.7**	…	**104.3**	**106.0**	**103.8**	**105.1**	**104.8**

* **(이전시점에 연결)** 스플라이싱은 시점 14에서 시작. 1-13시점 공표지수는 1차 산출에 의해 구한다. 시점 14 공표지수는 공표지수 시점 13지수에 1차 산출 시점 13과 k 변동과 2차 산출 시점 k(k=2,…,13)와 14 변동을 기하평균 적용하여 구한다 ;
$103.8\times[(100.7/103.8)(104.6/100.0)\times\ldots\times(103.8/103.8)(104.6/103.3)]^{1/12}=103.8\times1.012702=105.1$
시점 15 공표지수는 공표지수 시점 14지수에 2차 산출 시점 14과 k 변동과 3차 산출 시점 k(k=3,…,14)와 15 변동을 기하평균 적용하여 구한다 ;
$105.1\times[(100.2/104.6)(104.1/100.0)\times\ldots\times(104.6/104.6)(104.1/104.4)]^{1/12}=105.1\times0.996914=104.8$

* **(공표지수에 연결)** 스플라이싱은 시점 14에서 시작. 1-13시점 공표지수는 1차 산출에 의해 구한다. 시점 14 공표지수는 공표지수 시점 k 지수(k=2,…,13)에 2차 산출 시점 k와 14 변동을 기하평균 적용하여 구한다 ; $[(100.7\times(104.6/100.0)\times\ldots\times103.8\times(104.6/103.3)]^{1/12}=105.1$
시점 15 공표지수는 공표지수 시점 k 지수(k=3,…,14)에 3차 산출 시점 k와 15의 변동을 기하평균 적용하여 구한다; $[(100.6\times(104.1/100.0)\times\ldots\times105.1\times(104.1/104.4)]^{1/12}=104.8$

표4 13개월 롤링 윈도우, half splice : t-(T+1)/2+1이 연결 시점

시점	1	2	3	…	7	8	9	…	13	14	15
1 13	100.0	100.7	100.6	…	104.3	102.9	104.2	…	103.8		
2 14		100.0	100.2	…	103.8	102.5	103.7	…	103.3	104.6	
3 15			100.0	…	103.7	102.1	103.6	…	103.2	104.4	104.1
공표지수(이전시점 연결)	**100.0**	**100.7**	**100.6**	**…**	**104.3**	**102.9**	**104.2**	**…**	**103.8**	**105.0**	**104.6**
공표지수(기표지수 연결)	**100.0**	**100.7**	**100.6**	**…**	**104.3**	**102.9**	**104.2**	**…**	**103.8**	**105.0**	**104.7**

* **(이전계산에 연결)** 스플라이싱은 시점 14에서 시작. 1-13시점 공표지수는 1차 산출에 의해 구한다. 시점 14 공표지수는 1차 산출 시점 13과 8 변동과 2차 산출 시점 8과 14 변동을 공표지수 시점 13지수에 적용하여 구한다(103.8×(102.9/103.8)×(104.6/102.5)=105.0). 시점 15 공표지수는 2차 산출 시점 14와 9의 변동과 3차 산출 시점 9와 15의 변동을 공표지수 시점 14 지수에 적용하여 구한다(105.0×(103.7/104.6)×(104.1/103.6)=104.6).

* **(공표지수에 연결)** 스플라이싱은 시점 14에서 시작. 1-13기 공표지수는 1차 산출에 의해 구한다. 시점 14의 공표지수는 2차 산출 시점 8과 14의 변동을 공표지수 시점 8지수에 적용하여 구한다(102.9 × 104.6/102.5 = 105.0). 시점 15의 공표지수는 3차 산출 시점 9와 15의 변동을 공표지수 시점 9지수에 적용하여 구한다(104.2× 104.1/103.6 = 104.7).

표5 고정기준 : 이전 12월이 연결 시점, 시점 12(base 시점)

시점	1	2	3	4	5	…	11	12	13	14	15
1 13	100.0	100.7	100.6	101.6	102.7	…	104.3	106.0	103.8		
2 14		100.0	100.2	101.1	102.2	…	103.8	105.5	103.3	104.6	
3 15			100.0	101.0	102.0	…	103.5	105.3	103.2	104.4	104.1
공표지수	**100.0**	**100.7**	**100.6**	**101.6**	**102.7**	**…**	**104.3**	**106.0**	**103.8**	**105.1**	**104.8**

* **(공표지수에 연결)** 스플라이싱은 시점 14에서 시작. 1-13시점 공표지수는 1차 산출에 의해 구한다. 시점 14 공표지수는 2차 산출 시점 12와 14의 변동을 공표지수 시점 12지수에 적용하여 구한다(106.0×104.6/105.5 = 105.1). 시점 15 공표지수는 3차 산출 시점 12와 15 변동을 공표지수 시점 12지수에 적용하여 구한다(106.0 × 104.1/105.3 = 104.8).

시간 창과 스플라이싱 선택도 고려해야 한다. 스플라이싱 방법은 풀타임 윈도우에 걸쳐 산출된 지수와 비교하여 경험적으로 테스트 될 수 있다. 예를 들어, 몇 년에 걸친 데이터를 고려해 보자. 다변지수는 이 전체 기간에 걸쳐 계산될 수 있다. 이 지수는 완전히 이행성을 갖는 벤치마크로서 역할을 한다. 동일한 데이터로, 임의의 스플라이싱 기법을 사용하여 실시간 지수를 계산할 수 있다. 이 두 지수 사이의 차이는 스플라이싱 방법의 성능을 평가하는 데 도움이 될 수 있다. 그러나 '벤치마크' 지수

는 매우 긴 시간대의 영향을 받을 수 있다. 서로 다른 방법은 장단점을 가지고 있다. 고정기준은 전년 12월이 가격 기준기간으로 작용하는 양변 지수와 더 일치한다. 이동스플라이스는 가장 쉽게 이해하고 설명할 수 있는 방법이다. 다변지수로 얻은 월별 움직임과 일치한다. 그러나, 이 방법으로 연쇄 편의를 배제할 수 없다. 윈도우 또는 하프 스플라이스도 마찬가지이다. 평균 스플라이스는 가능한 모든 링크를 기반으로 하기 때문에 대안이 될 수는 있다. 또 다른 전략은 이전에 공표된 지수에 연결하는 것이다. 예를 들어 25개월 rolling time window, 하프 스플라이스(공표지수에 연결) 방법에 따르면 발표된 지수의 연간 변화율은 최근 계산된 다변지수의 연간 변화율에 해당한다. 룩셈부르크, 노르웨이, 벨기에가 적용하고 있고, 한국 통계청도 향후 다변지수를 고려하다면 이 선택을 우선 검토해 볼 필요가 있을 것이다.

5.2.5. [사례][16] 최근 룩셈부르크 변화 : 동적 basket → 다변지수 방법

'18.1월부터 통계청은 여러 소매업체로부터 받은 스캐너 데이터를 CPI에 포함(20년 CPI 바스켓 약 5% 차지)했고, 현재(21년) 스캐너 데이터 범위는 신선과일 및 채소를 제외한 COICOP 01(식품 및 비알코올 음료) 제품이다.

'18-'20년 스캐너 지수는 "동적 바스켓"(월간 체인 제본스) 방법을 사용하였고, 지수 산출에 모든 제품이 포함되는 것이 아니다.

- 특이치 없음(이상치 필터에서 제외)
- 2개월 연속 사용 가능
- 가장 많이 팔린 품목(컷오프별로 정렬)

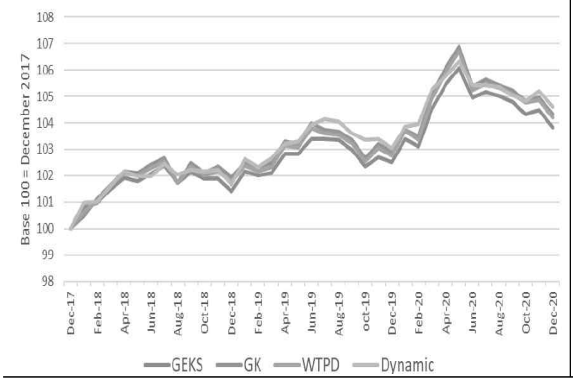
단점으로 덤핑 및 특이치 필터 사용, 샘플링으로 사용 가능한 모든 제품을 지수 계산에 통합할 수 없고, 매출액 정보를 지수 계산에 직접 반영할 수 없다는 것이고, 해결책으로 다변지수를 사용하는 것이다. 다변지수 방법 GEKS-Tq, GK, Weighted Time Product Dummy와 지수 수정 문제를 해결하기 위한 롤링 타임 윈도우 방식과 스플라이싱 방법을 검토하였다.

16) Recent developments in Luxembourgish CPI: from dynamic basket to multilateral methods, STATEC, Luxembourg

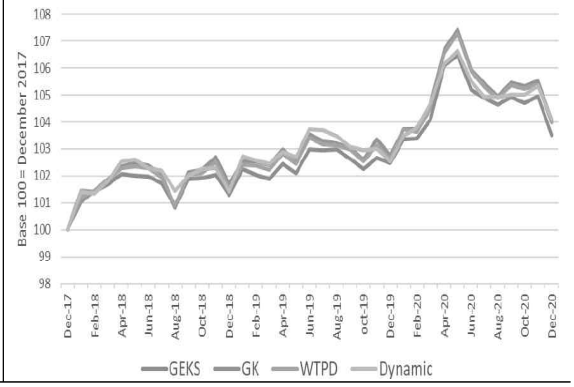
검토 결과 : Full Window 비교

“동적 바스켓“ 방법은 물가지수를 편향시키지 않는 것처럼 보이고, 이는 지수 방법 변화로 인한 영향을 많이 받지 않고 선택된 다변 지수 방법 중 하나로 대체 가능성이 있음을 나타낸다.

Without seasonal products



With seasonal products



검토 결과 : Splicing Methods 비교

GEKS이 GK 및 WTPD보다 full time window와 splicing 지수를 비교하여 변동성이 상대적으로 적고, full time window는 연쇄편의 없이 이행되므로, 이에 가까운 splicing이 일반적으로 선호된다고 할 수 있다.

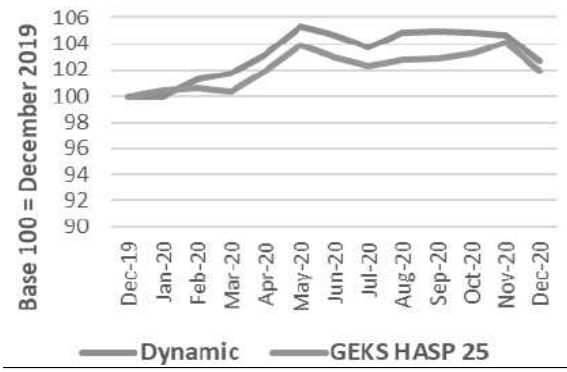
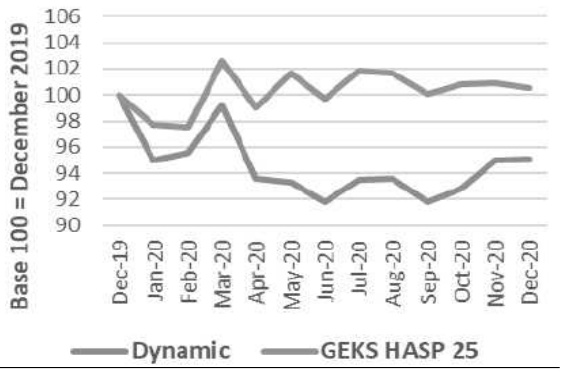
GEKS Without seasonal products

GK Without seasonal products

WTPD Without seasonal products

검토 결과 : 신구 방법 비교

GEKS은 GK 및 WTPD에 비해 splicing 선택의 변동성이 상대적으로 적고 연간변동률이 일치하는 가격지수를 생성할 수 있기에 25개월 기간 설정에 대한 “GEKS HASP“가 가격지수의 새로운 방법으로 선택되었다. 또한 결과는 25개월 물가지수 방식의 GEKS HASP가 동적 바스켓 물가지수 방식보다 계절상품 취급에 더 적합하다는 것을 보여준다.

신구 지수 비교(쌀)	신구 지수 비교(쵸코렛)
	

5.2.6. 다변지수 방법 평가

현재 소수 통계청만이 다변지수를 구현했고, 새로운 데이터 출처와 방법을 실행하려면 통계 영향, 유익성과 비용을 신중하게 고려해야 한다. 평가 프레임워크로 다음과 같은 7가지 통계품질을 고려 할 수 있다[17].

- 제도 환경(Institutional environment)—통계 생산자가 운영하는 제도 환경
- 관련성(Relevance)—통계가 사용자 요구를 얼마나 잘 충족하는지
- 적시성(Timeliness)—통계가 얼마나 빠르고 자주 공표되는지
- 정확도(Accuracy)—통계가 원하는 개념을 얼마나 잘 측정하는지
- 일관성(Coherence)—통계가 관련 정보 출처와 얼마나 일치하는지 여부
- 해석 가능성(Interpretability)—통계 통찰력을 제공하는 데 이용 가능한 정보
- 접근성(Accessibility)—통계 접근 용이성

주목 할 것은 다변지수가 표준 양변지수보다 더 복잡하여 현재 통계청의

17) ILO(2020). “Consumer Price Index Manual: Theory and Practice.” 10. Scanner data.

과제이다. 발표(생산)된 통계에 대해 사용 방법을 설명하고 투명성에 높은 가치를 두어야 한다.[18] 해석 가능성의 두 가지 측면, 첫째, 방법 자체가 지수 실무자와 사용자가 어느 정도까지 이해하기 쉬운지, 둘째, 각 지수가 생성하는 가격 움직임, 특히 이러한 움직임에 가장 큰 영향을 미치는 제품과 그 이유를 이해하기 쉬운지 여부를 고려해야 한다.

지수 공식을 선택시에 고려할 수 있는 여러 기준이 있고, 그 방법에는 장점과 한계가 있다. 일부기준이 아래 표에 요약되어 있다. 이러한 기준 외에도 실질적인 고려사항도 한 역할[19]을 할 수 있다.

다변지수 산출 방법 비교[20]

	GEKS	**WTPD**	**GK**
주요 원리	양변지수와 밀접한 관련, 물가지수 경제적 접근법과 가장 일치하는 방법	가격이 모델에 따라 생성된다고 가정, 가격지수에 대한 확률적 접근법에 기초	동 제품 수준 단위값 계산에 가장 가까운 가법 방법
테스트			
수량 동질성	만족	만족	실패
바스켓 테스트	실패	실패	만족
반응성 검정	만족	실패	실패
유연성	헤도닉 사용하여 누락된 가격대체 가능	회귀 분석에 제품 특성 포함 가능	헤도닉 사용하여 품질조정 단위 값 지수 추정 가능
양변지수 대응	가격비율 가중 기하평균, 가중치는 비중 산술평균(Törnqvist)	가격비율 가중 기하평균, 가중치는 비중의 조화 평균	바스켓 지수, 가중치는 수량의 조화 평균
민감성/견고성			
Splicing 선택	덜 민감	더 민감	더 민감
재고정리가격	더 민감	덜 민감	덜 민감

앞에서 언급했듯이 수정 불가능 지수를 얻기 위해 시간 윈도우 길이와 스플라이싱 기법을 결정해야 한다. 경험 관점에서, 일부 연구는 WTPD와

18) (Eurostat) 다변지수는 양변지수보다 더 복잡하다. 방법을 평가할 때 고려될 수 있는 한 가지 측면은 사용자에게 방법을 설명할 필요성이다. 어떤 방법이 설명하기에 다소 쉬운지를 평가하는 데는 주관적인 요소가 있다. 또한 사용자의 유형에 따라 다르다. 고급 사용자는 방법에 대한 보다 엄격한 설명에 관심이 있는 반면, 다른 사용자는 더 간단한 용어로 방법에 대한 설명을 받는 것을 선호할 수 있다. 세 가지 다변지수에 대해 다음과 같이 말할 수 있다. GEKS는 기본이 되는 양변 지수(예: Törnqvist 지수)를 고려하여 설명할 수 있다. WTPD는 기본 회귀분석을 통해 설명할 수 있다. GK는 품질조정 단위 값 지수로 설명될 수 있다. 방법의 설명 가능성 외에도, 기여도를 계산함으로써 해석 가능성을 촉진할 수 있다.

19) 예를 들어 GEKS보다 GK 또는 WTPD를 계산하는 프로그램을 실행하는 것이 시간이 더 많이 걸린다. 이 문제는 매우 큰 데이터 집합에서 문제가 될 수 있다.

20) Guide on the use of multilateral methods in the HICP Draft version (October 2021),

GK가 GEKS-Tq보다 이러한 선택에 더 민감하다는 것(더 달라진다)을 보여준다. 재고정리가격(덤핑)은 상품을 할인된 가격에 판매한 후 상품 분류(assortment)에서 제외될 때 발생한다[21]. 이 제품을 사용할 수 있는 마지막 달의 판매수량은 보통 적다. GEKS-Tq는 덤핑제품의 가격 하락에 더 큰 비중을 두는데, 이는 가격 하락이 과거 기간 해당 제품의 (더 큰) 지출 점유율에 의해 암묵적으로 영향을 받기 때문이다. 따라서 덤핑 필터는 GEKS-Tq와 함께 가장 잘 사용해야 한다.

12가지 시험 접근(test approach)[22]은 아래와 같다.

Test 1 : 이행성(transitivity) :
다변지수는 이행성을 만족한다. 기준기간 선택과 무관하며 연쇄편의를 피하는 것이 특성이다. HICP에서는 이행성을 만족하는 지수 공식이 허용된다고 명시적으로 언급하고 있다. $I^{0,t2}(P,Q) = I^{0,t1}(P,Q) * I^{t1,t2}(P,Q)$

Test 2: 동일성(Identity)
이는 모든 가격이 초기 수준으로 되돌아갈 경우 지수가 동일해야 한다. 모든 i=1..n에 대하여 $p_i^t = p_i^0$으로 하자. 그러면, $I^{0,t}(P,Q) = 1$

Test 3: 여러 시점 동일성(multi period identity test)
가격, 수량 마지막이 첫 번째 시점과 동일할 경우, 연결된 지수가 마지막에 다시 동일한 상태로 되돌아갈 것이 요구된다. 모든 i=1.n에 대해 $p_i^T = p_i^0 \text{ and } q_i^T = q_i^0$ 인 경우, $I^{0,t}(P,Q) * I^{1,2}(P,Q) * \cdots * I^{T-1,T}(P,Q) = 1$

Test 4: 연속성, 양성 및 정규화(Continuity, positivity and normalisation)
$I^{0,t}(P,Q)$는 가격, 수량 자료의 양의 연속 함수이고 $I^{0,0}(P,Q)$은 1이다.

Test 5 : 비례 가격 검정(Proportional prices test)
모든 기간 가격이 비례하는 경우, 물가지수는 이러한 비율에만 의존한다. $p_i^t = a^t p_i^0$ 와 같은 a^t이 모든 i=1..n, t=0,..,T에 대해 있다고 가정하면,

---

21) 가격이 비정상적이고 수량이 적은 제품(재고정리 가격)에 대한 GEKS-Törnqvist 민감성을 개선하기 위해 ABS(호주 통계청)는 필터를 사용하여 이러한 제품을 지수 산출에서 제거한다. 재고 정리 가격인 제품을 제외하는 것은 CPI 현재 관행과 일치한다

22) Guide on the use of multilateral methods in the HICP Draft version (October 2021),

$I^{0,t}(P,Q) = \alpha^t \forall\, t = 0 \ldots T$

Test 6: 수량 동질성(Homogeneity in quantities)
어느 한 시점에서 수량을 재조정한다고 물가지수가 달라지는 것은 아니다. 모든 i=1..n에 대해 어느 시점 k를 위해 $\hat{q}_i^k = \gamma q_i^k$와 같은 γ가 존재한다면,
$I^{0,t}(p^0,\ldots,p^k,\ldots,p^T,q^0,\ldots,q^k,\ldots q^T) = I^{0,t}(p^0,\ldots,p^k,\ldots,p^T,q^0,\ldots,\hat{q}^k,\ldots,q^T) \forall\, t = 0 \ldots T$

Test 7: 가격 동질성(Homogeneity in prices)
어느 한 시점의 가격을 다시 조정하면 물가지수가 같은 비율로 변경된다. 모든 i=1…n에 대하여 $k \neq 0$ 시점에 대하여 $\hat{p}_i^k = \gamma p_i^k$이 존재한다면,
$\gamma I^{0,k}(p^0,\ldots,p^k,\ldots,p^T,q^0,\ldots,q^k,\ldots q^T) = I^{0,k}(p^0,\ldots,\hat{p}^k,\ldots,p^T,q^0,\ldots,q^k,\ldots,q^T)$

Test 8 : 같은 단위로 측정(Commensurability)
가격과 수량 표현 단위를 바꾼다고 해서 물가지수가 바뀌는 것은 아니다.
$Let \delta_i > 0 (i = 1 \ldots n).\ Let\ \hat{p}_i^t = \delta_i p_i^t$ 그리고 $\hat{q}_i^t = \frac{q_i^t}{\delta_i} (i = 1 \ldots n \text{ and } t = 0 \ldots T)$ 이면,
$I^{0,t}(p^0,\ldots,p^T,q^0,\ldots,q^T) = I^{0,t}(\hat{p}^0,\ldots,\hat{p}^T,\hat{q}^0,\ldots,\hat{q}^T) \forall\, t = 0 \ldots T$

Test 9: 시점 처리의 대칭성(Symmetry)
시점을 다시 정렬해도 시점 간 가격 변동은 변경되지 않는다.

Test 10: 상품 처리 대칭성
상품 순서를 다시 지정해도 가격 지수는 변경되지 않는다.

Test 11: 가격 바스켓 테스트(basket test for prices)
기준과 현재 시점 수량이 동일하면 현재 시점 지수는 바스켓 지수에 해당한다. 모든 i=1..n에 대해 $q_i^0 = q_i^k = q_i$이면, 다음과 같다.

$$I^{o,k}(P,Q) = \frac{\sum_i p_i^k q_i}{\sum_i p_i^o q_i}$$

Test 12 : 대체가격 반응성 검정(responsiveness test to imputed prices) 하나 이상의 시점에 누락된 제품이 있으면 이러한 누락 제품에 대한 대체 가격이 있고 해당 제품의 수량은 0으로 설정된다. 반응성 검정에 따르면, 다른 가격 지수를 제공하는 누락 가격의 다른 대체가 존재해야 한다.

다음 표는 이러한 테스트와 관련 세 다변지수의 특성을 요약한 것이다. 이것은 단지 선택된 테스트일 뿐이며, 어떤 특성이 어떤 다변지수 공식에 의해 충족되는지 명확히 하기 위해 다른 테스트가 정의될 수 있다.

다변지수 테스트 결과

		GEKS-Tq	WTPD	GK
Test 1	이행성(transitivity)	Yes	Yes	Yes
Test 2	동일성(Identity)	No	No	No
Test 3	여러 시점 동일성(multi period identity test)	Yes	Yes	Yes
Test 4	연속성, 양성 및 정규화	Yes	Yes	Yes
Test 5	비례 가격 검정(Proportional prices test)	Yes	Yes	Yes
Test 6	수량 동질성(Homogeneity in quantities)	Yes	Yes	No
Test 7	가격 동질성(Homogeneity in prices)	Yes	Yes	Yes
Test 8	같은 단위로 측정(Commensurability)	Yes	Yes	Yes
Test 9	시점 처리 대칭성	Yes	Yes	Yes
Test 10	상품 처리 대칭성	Yes	Yes	Yes
Test 11	가격 basket 테스트	No	No	Yes
Test 12	대체가격 반응성 검정	Yes	No	No

세 가지 다변 지수 모두 테스트 1, 3, 4, 5, 7, 8, 9 및 10을 만족하며 검정 2를 만족하는 지수는 없다. 테스트 2의 실패는 동일성 시험을 만족하는 양변지수 방법과 비교하여 다변지수의 한계로 볼 수 있다. GK의 주요 특징은 테스트 6을 충족하지 못한다. GK는 각 기간 판매수량에 따라 달라서, 제품이 더 많이 팔리는 달은 종종 결과에 더 큰 영향을 미친다. 이는 매월 동일한 방식으로 처리하는 GEKS-Tq 또는 WTPD 경우에는 해당되지 않는다. 마지막으로, GEKS-Tq 특징은 누락제품의 대체가격을 통합할 수 있다(테스트12). GEKS-Tq의 표준적용에서 가격은 대체되지 않고, time window의 두 기간 동안 일치하는 제품만 고려된다. 이 테스트에서는 해당 기간 동안 제품이 판매되지 않았더라도 예상 가격이 고려될 수 있고, 가격추정은 헤도닉을 사용하여 수행될 수 있다.

5.2.7. 추가 고려 사항

스캐너 자료는 통계청 활용 데이터 출처와 방법에 상당한 변화이며, CPI 사용자와 이해관계자에게 다음과 같은 활동으로 잘 전달되어야 한다.

- 새로운 방법 및 데이터 출처 관련 정보 공개
- 주요 이해관계자(예: 한국은행, 기재부 등) 및 기타 이해관계자(예: 학계, 공공부문 포함)와 대면 회의 수행
- 공개적으로 알리는 데 미디어 및 기자 브리핑 사용
- 이해관계자 및 대중이 통계청에 의견을 제출하도록 권장
- 변경사항 검토 및 지원을 장려하기 위해 관련 학계(학자)들과 협력

몇 년이 걸릴 수 있는 이 협의에, 통계청은 협의 과정에서 제기된 사항에 대응하고, 이 접근법을 뒷받침하는 근거와 경험적 결과를 포함하여 CPI를 산출하기 위해 스캐너 데이터를 어떻게 사용할 것인지 설명하는 내용을 공개하고, 데이터 출처와 방법을 명확히 기술해야 하며 변경사항 진행에 대한 시간표(timetable)를 제공하는 것이 바람직하다.

변경 관련 공개를 한 후에는, 통계청은 현행과 새로운 데이터 출처와 방법을 병행하여 약 6개월 동안 CPI를 산출할 것이 제안된다. 이 전환기간을 통해 CPI를 산출하기 위한 프로세스를 개선하고 두 접근의 결과를 비교할 수 있다. 이 기간은 통계청이 CPI처리 및 계획된 일정에 따라서 새로운 데이터 출처와 방법을 실시간으로 사용할 수 있는 첫 번째 기회일 수 있다. 병행처리 결과 공개여부는 통계청 재량이다. CPI산출에 새로운 데이터 출처와 방법을 적용하는 첫 번째 기간은 사전에 잘 공개되어야 하며 미디어 및 기타 주요 데이터 사용자를 위한 상세한 메타데이터를 포함해야 한다. 이것이 CPI에 구현된 방법론적 변화를 잘 이해되도록 보장할 것이다.

6. 영국 활용 사례

6.1. 스캐너 데이터 사용에 관한 연구[23)]

영국 CPI는 2023년부터 스캐너 데이터, 웹 스크래핑을 기존 데이터 수집 방법과 통합을 계획하고 있다. 현재 8개 소매업체로부터 스캐너 자료를 받고 있고, 이 데이터를 실제 생산하는 시점에 시장 범위 확대를 위해 더 많은 소매업체와 협의 중이다. 지금까지 주요 초점은 식료품이다. 또한 이 데이터 처리하는 방법이 제품 분류, 시간범위, 고유 및 재출시 제품 식별·추적 방법, 그리고 제품 크기 변화를 고려하면서 가격 도출, 할인 처리 등을 포함하여 상점에서 대면으로 수집하는 데이터 처리와 어떻게 다른지 연구해 왔다,

6.1.1. 데이터 수집 및 품질 확보

식료품은 시장의 많은 부분을 적은 수의 소매업체들로 다룰 수 있어서, 현재 식료품 스캐너 자료 수집을 진행해왔다. 2023년부터 시행은 주로 식품, 음료 및 담배 제품에 초점을 맞추고 있지만, 향후 범위를 확대할 계획이다. 주요 소매업체와 자료 수집에 대해 협력하고 있으며, 각 업체마다 수억 개의 데이터, 처리해야 하는 양이 크게 증가하고 있다.

정기적 프로세스로 획득한 데이터는 품질검사를 거친다. 초기검사에는 변수가 지정된 유형이고 값이 미리 정의된 범위 내에 있는지, 자료 크기와 모양이 예상대로인지 확인하는 작업이 포함된다. 초기점검이 실패하면 소매업체의 데이터 재전송 문제가 발생한다. 일단 이러한 초기검사를 통과하면, 새로운 품목조사, 이상값 식별 등 추가검사가 이루어진다. 현재 이러한 새로운 자료를 실제 생산에 사용할 때 품질표준을 충족하는지 확인하기 위해 검토를 진행하고 있다. 이 작업에는 새로운 데이터 출처와 관련된 위험 식별 및 취할 수 있는 조치를 포함하여 데이터 품질 보증을 업데이트하는 데 필요한 정보 수집을 하고 있다.

---

23) ONS, Research into the use of scanner data for constructing UK consumer price statistics April 2021

6.1.2. 스캐너 데이터 처리 방법

지수(테스트)에 대한 여러 처리 결정이 미치는 영향을 검토하기 위해 단일 소매점으로부터 스캐너 데이터를 사용할 수 있는 허가를 받아서 13개월 이동 splice 적용 GEKS-Törnqvist을 사용했다.

식품, 음료 및 담배 제품에 대한 스캐너 데이터 분류

데이터를 받고 충분한 품질을 가지고 있다면, 프로세스 첫 번째 단계 중 하나는 데이터를 적절한 범주로 분류하는 것이다. 예를 들어, 체다 덩어리를 “체다 치즈“로 분류하거나 핑크 레이디 사과 한 봉지를 "사과" 로 분류하는 것이다. 분류 방법에 대한 이전 연구는 주로 machine learning을 사용하여 웹 스크래핑 의류를 범주(“여성 청바지“, “아기 파자마“ 및 “남자 셔츠“)로 분류하는 자동화 방식에 초점을 맞추어 왔다. 의류에는 자료 양 때문에 대부분 자동화 접근이 필요하지만, 식료품은 자료 속성이 다르기 때문에 다른 방법이 더 적합할 수 있다.

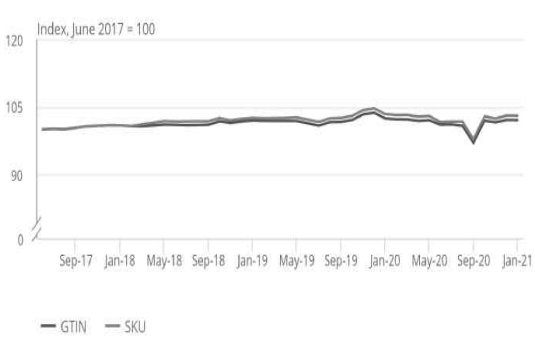
시간 범위

현재 대부분의 소매업체로부터 주간 단위로, 일부는 일일 자료를 받고 있지만, 주간 단위 집계 데이터인 경우, 일부 주가 인접한 두 달에 속하고, 한 주 내 거래가 속하는 달을 결정하기 위해 데이터를 세분화할 수 없다. 가격은 해당 월에 귀속되어야 월 사이 발생하는 차이가 확인될 수 있다. HICP 스캐너 데이터 가이드(2017)에서는 “가능한 한 많은 날짜를 포함해야 하고, 다른 달을 참조하는 자료는 포함하지 않아야 한다. 시간 간격이 연중 동일한 방식으로 정의되도록 하는 것이 중요하다.“ 고 명시하고 있다. 먼저 데이터의 부분 월을 사용하여 편향을 평가할 계획이고, 편향이 발생하는 것으로 나타나는 경우, 어떤 옵션(예 : 2주 vs 3주 vs 전체 월)이 편향을 최소화하는지 살펴볼 것이다.

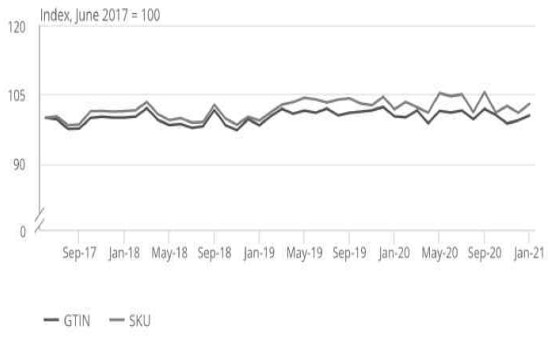
스캐너 자료 고유 제품 식별

제품 포장이나 크기 변화가 있을 경우, 재출시되는 경향이 있다. 가격지수 산출시, 재출시된 제품이 기존제품과 유사하다면, 이 제품을 계속 추적한다. 재출시는 종종 가격 또는 작은 품질 변화와 관련된다. 이에, CPI에 미치는 영향을 파악할 수 있도록 그것들을 포착하고 품질 변화에 맞게 조정하는 것이 중요하다. 모든 스캐너 데이터에는 제품을 식별하는 데 사용할 수 있는 상품품목번호 GTIN과 소매업체 정의 SKU코드가 포함되어 있다. GTIN은 각 제품에 고유하며 소매점마다 일관되나, 포장과 같은 작은 변형이나 재료의 사소한 변경으로 인해 변경될 수 있다. SKU는 소매업체에서 정의한 코드이지만 일반적으로 GTIN보다 넓기 때문에 가격지수를 산출하는 데 더 적합한 제품을 식별할 수 있다. SKU는 포장, 크기 또는 성분에 사소한 변화가 있을 경우 변경되지 않지만, 신 GTIN 코드가 동반된다. 그러나 재출시가 SKU에 의해 포착된다는 보장은 없다. SKU는 소매업체가 정의한 것이며, 한 소매업체는 SKU내 재출시를 포착하지만, 다른 업체는 SKU를 GTIN과 같이 보조를 맞춘다. 다른 국가통계기관은 제품 설명에 대한 텍스트마이닝, 제품특성 매칭, 지출동향 감지 등과 같은 기법을 사용하여 이전제품과 새로 재출시된 제품의 매칭을 시도한다. 이런 편향이 데이터에 존재하는지를 테스트하기 위해 SKU와 GTIN를 고유제품 식별자로 사용하여 가격지수를 작성하여 비교하였다 (그림 1에서 4). 이 비교는 종종 기초집계 수준에서 지수 값 차이를 보이지만, 어떤 방향으로 명백한 편향이 없다.

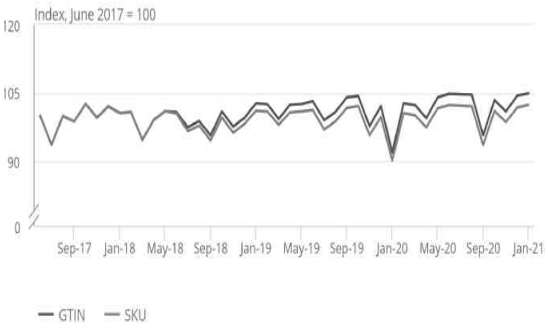
< 단일 소매점 : 고유 제품 식별자 GTIN vs SKU 사용할 경우의 영향>



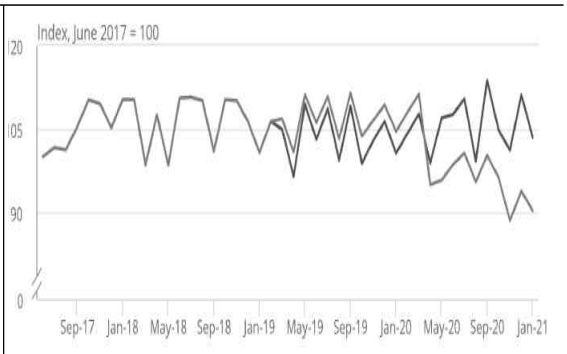
1. salad bags



2. children's sweets



3. children's chocolate



4. frozen peas

데이터의 정밀검토와 제품 재출시에 대한 초기조사를 통해 소매업체의 동일한 SKU에서 모든 제품 재출시를 적절히 포착하지는 못하는 것 같다고 판단된다. 그래서 제품 재출시 확인 및 연계를 위한 기록 연계 방법을 연구하고 있다. 이 방법은 기존 SKU 자료에서 각 새로운 SKU를 평가하는 것이다. 일정기간 내에 매칭이 확인되면 새 SKU가 기존 SKU에 자동으로 연결된다. 기존에 연동될 수 없는 새로운 SKU 각 제품에 대해 상품명과 가격의 유사성을 바탕으로 상품계층(범주) 내에서 기존 상품이 검색된다. 매칭에 대해 정량적으로 점수가 매겨지며 매칭율이 높은 제품은 검증을 위해 반환된다. 기존제품과 매칭율이 높은 제품이 자동으로 검증될 수 있는지, 아니면 수동검증이 필요한지 검토중이다. 신제품이 기존 제품과 매칭율이 낮으면 신제품으로 취급할 수 있다. 여전히 제품매칭을 검증하기 위한 다양한 임계값의 적합성을 조사하고 있다. 매칭율이 높고 중간 및 낮은 제품의 사례가 아래 표에 제시되어 있다.

< SKU 매칭 예 : 기록 연계 방법을 사용하여 SKU 기반 매칭 검토 >

Product with new SKU	Best match from all existing products	Match % on name and price	Match likelihood
BrandX Leek and Potato Mini Pies 4 Pack	BrandX Leek and Potato Small Pies 4 Pack	92.70%	Match
BrandY Mixed Fruit Energy Drink 250ml	BrandY Energy Drink 250ml	87.50%	Likely match
BrandZ Mixed Berry Mints 38G	BrandA Sugar-Free Strawberry Sweets 120G	26.20%	No match

불일치 측정 단위(UoM : units of measurement) **및 가격 도출**

제품 크기 또는 중량은 물가 측정에 사용되는데, 단일 SKU코드 제품도 시간이 지남에 따라 여러 개의 측정 단위(UoM)[24]를 가질 수 있고, 한 소매업체('17.6~ '19.6월)에서 복수 UoM을 갖는 SKU비중이 3.6%이기에 가능한 경우 UoM을 표준화하여 시간이 지남에 따라 일관성이 보장되도록 했다. 예를 들어, 킬로그램 또는 그램 단위로 표시된 UoM은 그램으로, 리터, 센티미터, 밀리리터로 표현된 UoM은 밀리리터로 표준화된다. 이런 방식으로 단위를 표준화하더라도 같은 기간 복수 UoM을 보유한 SKU는 여전히 지출의 2.8%를 차지했다. SKU는 하나의 SKU에 사용되는 모든 UoM을 공통 UoM으로 변환하는 것이 항상 가능한 것은 아니다. 예를 들어 제품이 그램 단위로 측정되지만 "팩"으로만 지칭된 경우에는 단위를 상호 변환할 수 없다. 이러한 이유로, 제품 SKU코드와 표준화된 UoM 조합이 지수 계산을 위해 제품을 매칭시킬 때 고유 식별자를 형성하기 위해 사용될 수 있다. 즉, 불일치 UoM을 사용하는 제품 간 비교를 피하기 위해 동일한 SKU를 사용하지만 UoM이 다른 제품은 고유 제품으로 취급한다.

스캐너 데이터는 제품의 명시적 가격이 포함되어 있지 않으며, 대신 각 고유제품에 대한 주간 지출(또는 매일)과 판매 수량을 제공한다. 이를 통해 각 제품의 총지출(V)을 판매수량(q)으로 나누어 평균가격(p)을 구할 수 있다. $p_i^t = \left(V_i^t\right) / \left(q_i^t\right)$, 이 식을 사용, 한 주에 100파운드(V) 사과(i)가 200개 사과가 판매되었다면, 사과당 50펜스의 평균가격을 도출할 수 있다. 제품 크기나 중량 변화에 따른 물가지수를 지동으로 계산하기 위해, 다음과 같은 변환이 이루어질 수 있다.

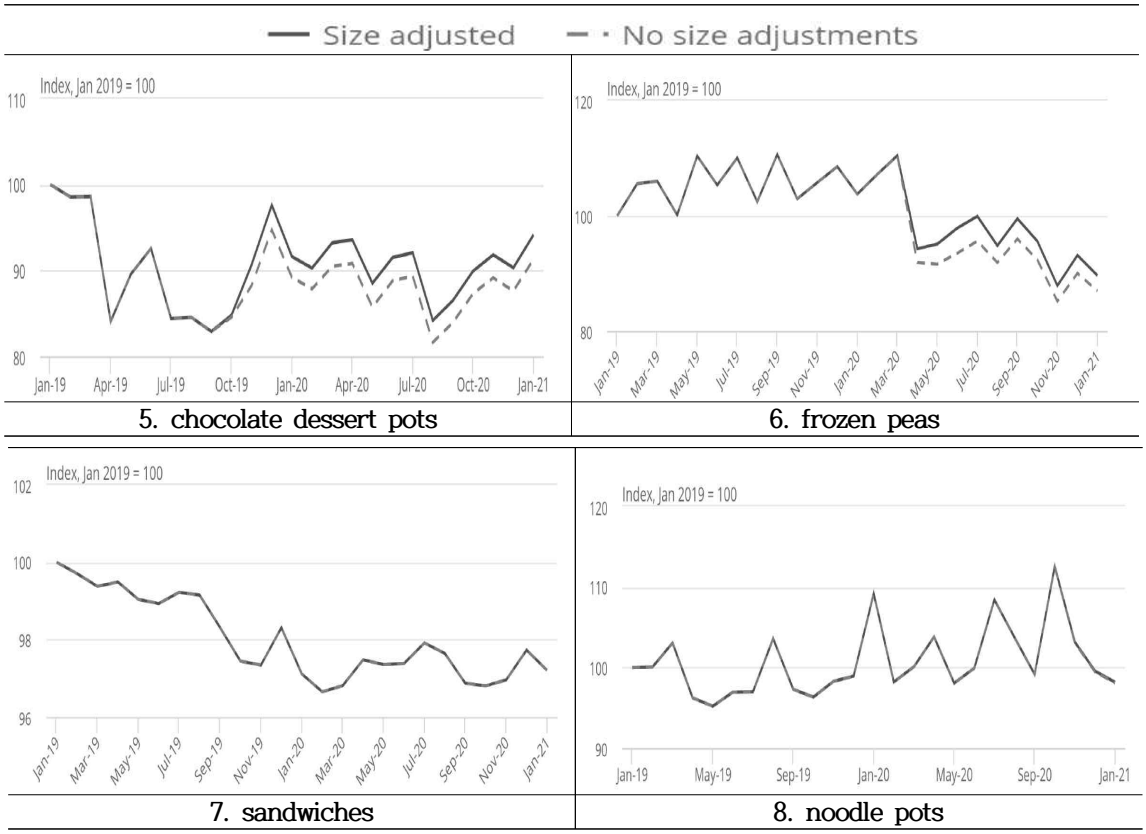
① 측정 단위당 가격을 산출하기 위해 총 지출을 판매수량, 제품크기 또는 중량으로 나눈다. ② 판매수량에 제품크기 또는 중량을 곱하여 구입 총 크기 또는 중량을 구한다.

스캐너 데이터를 사용하여 하위수준 집계에 크기를 조정할 경우의 영향을 검토했다. 많은 경우에 크기조정은 거의 또는 전혀 영향을 미치지 않지

24) 자료에 포함된 측정 단위(UoM) 예로는 그램, 킬로그램, 리터 및 packs 이 있음

만, 일부 경우 제품크기 변화를 고려할 때 인플레이션이 고조되는 것을 볼 수 있다. 그림 5와 6에서 크기가 조정된 집계는 조정되지 않은 집계에서 위쪽으로 갈라져 크기 감소가 이 기간 동안 인플레이션에 기여했고, 그림 7과 8에서 지수 값 간에는 차이가 별로 없어 이 기간 동안 이들 제품의 크기 변화 영향이 거의 없었음을 시사한다.

< 단일 소매점 : 하위 레벨 aggregate의 크기 변화에 대한 조정의 영향>



5. chocolate dessert pots | 6. frozen peas

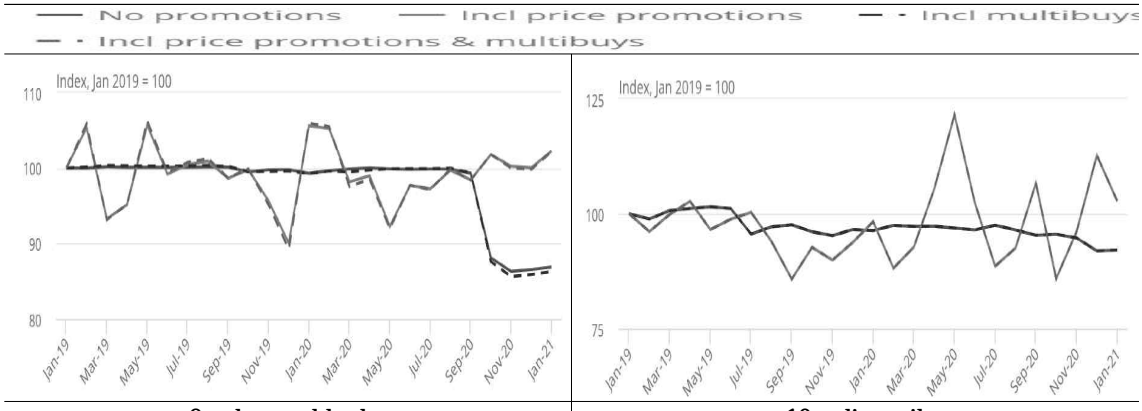
7. sandwiches | 8. noodle pots

스캐너 데이터 할인 처리

CPI에서는 가격 프로모션이 고려되지만, 다량구매(multibuy) 프로모션은 소비자 점유 비율을 알지 못하므로 현재 포함되지 않는다. 또한 같은 이유로 차별 할인(예: 로열티 카드 할인, 직원 또는 학생 할인)을 고려하지 않는다. clear (노란색 스티커) 제품으로 할인된 가격은 수집되지 않는다. 품질이 전체 또는 판촉 가격으로 판매되는 제품과 비교할 수 없는 것으로 고려되기 때문이다. 이 데이터에는 가격 인하, 멀티바이 할인 및 차별 할인을 포함한 모든 유형의 할인에 대한 정보를 갖고 있다. 즉, 할인변경 영향을 포함하여 소비자가 평균적

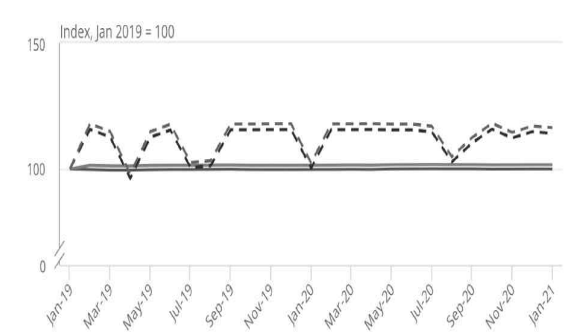
으로 지불하는 실제가격을 더 잘 반영할 수 있다. 단일 소매점 데이터(19.1~21.1월)를 사용하여 하위 수준 집계의 가격인하 및 다량구매 할인이 미치는 영향을 검토했다. 대부분의 하위 수준 집계에 있어, 프로모션 포함은 이들이 고려되지 않은 경우에 비해 상대적으로 변동성을 가져왔다. 예를 들어, 일부 제품의 경우, 2020년 COVID로 인해 전국적인 lockdown 기간에 가격이 상승했다고 했지만, 프로모션 영향이 배제되었을 때 이런 효과는 제거되었다(예:그림10 올리브오일). 일반적으로 가격 프로모션이 물가지수에서 제외된다면, 가격변동에 대한 소비자 경험을 적절히 포착하지 못한다. 이 단일 소매업체에서 관측된 대부분의 경우, 가격 프로모션은 다량구매 할인보다 가격지수에 더 큰 영향을 미쳤다. 그러나 이는 제품 범주에 따라 달랐다. 예를 들어, 치즈 블록과 올리브 오일 범주에서 가격 프로모션에 비해 다량구매 할인은 거의 영향을 미치지 않았으며(그림 9와 10), 할인은 베리 및 초콜릿 바 경우 가격 프로모션보다 큰 영향을 미쳤다(그림 11과 12). 이는 multibuy 할인이 제품별로 다르다는 것을 보여주지만, 이는 개별 소매점의 가격 전략에 따라 크게 달라질 수 있다.

< 단일 소매점 최소 단위 집계 : 가격 프로모션 및 다량구매(multibuy) 영향>

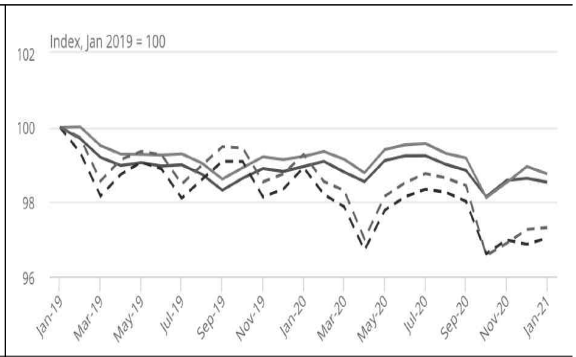


9. cheese blocks

10. olive oil



11. berries



12. chocolate bars

6.2. 적용 산식 연구[25)]

영국 ONS가 2020.9.1. 적용산식에 관련하여 진행 연구결과를 홈페이지에 수록하였고, 2021.6월 UNECE CPI 전문가회의에서 새로운 자료 소스로서 스캐너데이터 CPI 활용 관련하여 발표하였기에 이를 검토하였다.

영국통계청이 '23년에 사용할 데이터는 스캐너와 웹 스크래핑 자료이고, 기존 데이터에 비해 제품범위와 수집빈도가 증가하여 데이터 자동화 수집 방법을 연구하고 있다. 이러한 자료는 활용을 극대화하기 위해 새로운 지수산식이 필요할 수 있다. 그러나 현재 국제적인 합의가 거의 없이 사용할 수 있는 지수방법의 수는 다양하다. 이에 품질프레임워크를 만들어 이들 데이터 사용시, 다른 국가통계기관에서 사용 중인 다변지수를 포함하여 최저수준의 집계에서 CPI를 산출하는 데 사용할 수 있는 다양한 방법에 대한 테스트를 수행했다. 그 결과 최저 수준 집계에서 다변지수 사용이 고정 또는 연쇄 양변지수 방법을 사용하여 만들어진 지수보다 더 포괄적이고 정확할 것이라는 것을 보여주었다. 품질조정 Geary Khamis(QU-GK)는 지수방법 프레임워크 기준에 대해 가장 높은 점수를 얻은 방법이었고, 테스트 하에서 잘 수행되었다. 따라서, QU-GK 방법이 지출정보 또는 그 근사치를 이용할 수 있을 때 스캐너 및 웹 스크랩된 자료와 함께 사용할 현재 제안된 방법('21년 기준)이다.

6.2.1. 수행 단계(Timelines), 로드맵

'23년 CPI에 대체 데이터 통합을 위해 3단계 접근방식을 추진하고 있다.

1단계(연구, 개발 2020) 기존의 소스 및 방법과 함께 대체 데이터 소스를 함께 사용할 수 있는 시스템과 방법을 개발
2단계(적용, 개발 2021) 소비자 가격 관련 이해 관계자 자문패널 우선순위에 따라 우선 적용 품목 범주*에 방법 적용
* / / / / / / 도 요금 등
3단계(참여, 병행 2022) 대체 데이터가 CPI에 미치는 영향에 대한 분기별 실험 추정치 공개와 이런 변화에 대한 이해관계자 및 사용자의 참여

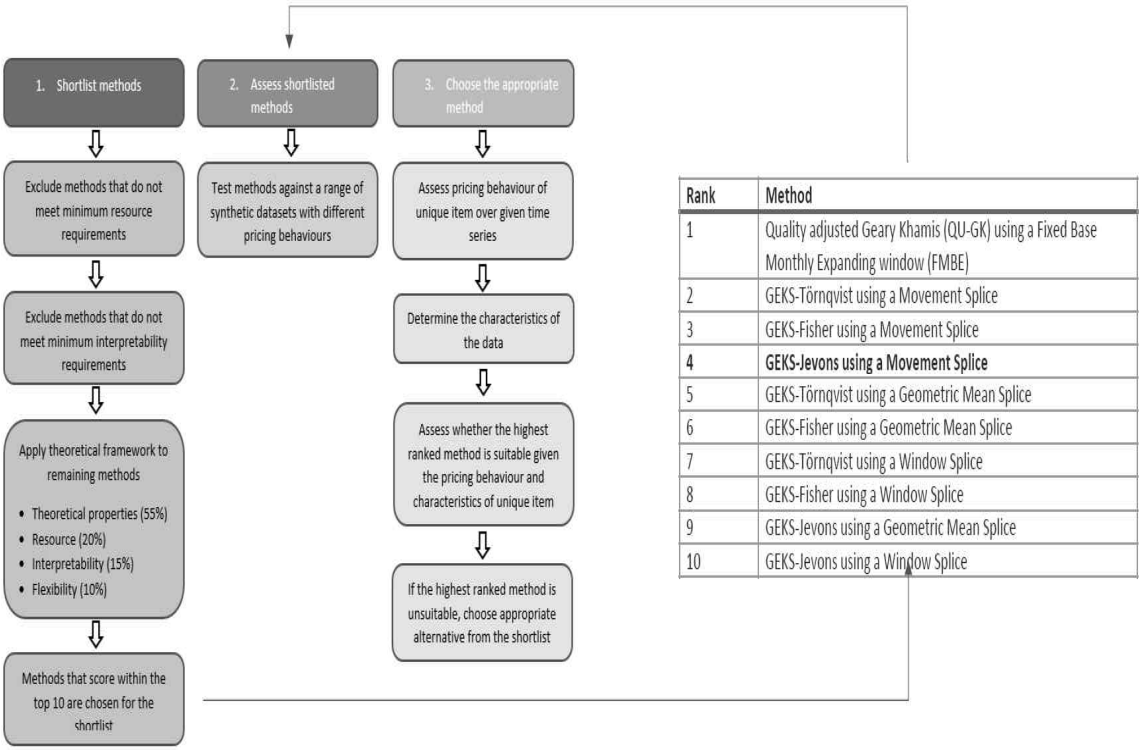
25) ONS, New index number methods in consumer price statistics 1 September 2020

6.2.2. 지수 산식 선택 과정

1단계(최종 후보 대상 선정) 미리 정해진 기준의 이론적 틀에 따라 점수를 매겨 상위 10위까지 우선순위가 매긴다. 많은 방법이 바람직하지 않지만, 단일 방법이 모든 데이터 출처 및 지출 범주에 적합하지 않을 수도 있다.

2단계(최종 후보 대상 평가): 어떤 방법이 어떤 시나리오에서 잘 작동하는지를 알아내기 위해 다른 가격 행위(예: 노후화, 높은 변동성)를 포함하는 여러 데이터에 대해 방법을 평가한다.

3단계(방법 선택) : 그런 다음, 상위 순위가 매겨진 방법이 효과가 있는지 확인하기 위해 여러 시장(예: 식료품, 의류)의 행동을 평가한다.



Rank	Method
1	Quality adjusted Geary Khamis (QU-GK) using a Fixed Base Monthly Expanding window (FMBE)
2	GEKS-Törnqvist using a Movement Splice
3	GEKS-Fisher using a Movement Splice
4	**GEKS-Jevons using a Movement Splice**
5	GEKS-Törnqvist using a Geometric Mean Splice
6	GEKS-Fisher using a Geometric Mean Splice
7	GEKS-Törnqvist using a Window Splice
8	GEKS-Fisher using a Window Splice
9	GEKS-Jevons using a Geometric Mean Splice
10	GEKS-Jevons using a Window Splice

다음은 지수방법을 선택하는데 중요하다고 생각한 기준들이다. 자문패널 및 소비자가격 이해관계자들과 이런 기준의 중요성을 논의했고, 중요성을 반영하기 위해 가중치를 적용했다. 각 방법별로 기준에 대해 평가 점수를 매기고, 그런 다음 점수를 가중 합산하여 각 방법의 최종 점수를 얻는다.

· 이론 특성 (Theoretical properties)

정확하고 신뢰할 수 있는 통계작성을 고려하여, 이론 특성은 55%라는 가장 높은 가중치이며, 세 가지 핵심 영역으로 나눌 수 있다.

① 공리/시험 접근법(30%), ② 이행성(15%), ③ 특성(10%)

공리/시험 접근은 이론 관점에서 방법 적합성을 측정하기 위한 것이고. 이론 특성이 논의될 때 종종 고려되는 두 가지 접근법, 즉 확률 접근과 경제적 접근이 있다. 현재 생활비 지수(COLI)를 고려하지 않는다는 사실 때문에 경제적 접근법은 가중치 0으로, 확률 접근도 고려됐지만 확률론에서 이루어진 분산가정이 가격 움직임과 일치하지 않는다는 비판에 기초, 가중치가 0으로 부여되었다. 이행성 부족은 정적 제품군에서는 체인이 필요하지는 않기에 크게 중요하지 않으나, 동적 제품군에서는 중요하고, 고빈도 연쇄는 높은 변동(이탈)을 설명할 수 있지만 상당한 연쇄편의를 초래할 수 있다. 특성은 먼 시기의 가격변동 영향이 최소화 되어야 한다. 예를 들어, 9분기 윈도우 다변지수는 2년 이상 전의 데이터를 필요로 한다. 지수는 그 달 내에 발생한 물가 변동의 특성을 더 잘 나타내야 하며 2년 전에 관찰된 물가 변동의 특성은 덜해야 한다.

· 자원 (Resource)

20% 가중치이며, 처리요구사항을 잘 관리할 수 있는가? 이다. 이 기준은 “이 방법이 인적 및 정보 자원을 보다 효과적으로 사용할 수 있게 하는가?“라는 질문에 답하는 것을 목표로 한다. 필요한 컴퓨팅 요건과 처리 능력에서, 양변지수는 지수 계산하는 데 더 적은 기간을 요구하기에 다변지수 보다 덜 집약적이다. 또한 지출자료 또는 제품 특성(헤도닉)을 사용하는 방법은 사용되는 변수 증가로 인해 정보자원에 더 큰 부담을 줄 것이다.

· 설명력 (Interpretability)

15% 가중치이며, 사용자는 그 방법을 이해하고 있는가? 가격변동이 쉽게 해석될 수 있는가? 이다. 통계기관은 생산하는 통계를 투명하고 발표된 통계에서 사용되는 방법을 정당화하는 것이 필수적이다.

· 유연성 (Flexibility)

10% 가중치이며, 방법들이 다양한 목적, 데이터 출처 및 품목 유형을 위해 어떻게 사용될 수 있는지를 평가한다.

· 응집 (Cohesion)

내부 결속(cohesion)은 바스켓의 서로 다른 영역과 다른 데이터 소스 관련이고, 외부 결속은 국가 간 비교 다른 NSO와의 관련이다. 사용자 피드백이 주어지면 응집 기준은 방법에 점수를 매긴 다기 보다는 동일한 점수을 분리하기 위해 이차필터로 사용할 수 있다. 이는 단순히 다른 곳에서 그 방법을 사용하고 있기 때문이 아니라, 그 방법이 품질을 고려하여 선택되어야 하기 때문이다.

다변지수를 사용하면 제품 적용범위가 넓어지고 제품 수준에서 가중치를 매길 수 있으며 동적 데이터에서 수집된 제품 정보를 더 잘 활용할 수 있다. 추가적인 결과는 다음 시뮬레이션에서 살펴보고자 한다,

7. 시뮬레이션

7.1. Eurostat[26)]

공개되있는 Dominick 스캐너자료를 활용(88개월, '90.1~'97.4월)하여 주간 데이터를 월별로 변환, 데이터는 체인 수준까지 모든 상점에서 집계, 개별 제품 정의 품목코드가 사용되었고, 다음 6가지 제품에 초점을 맞추었다.

- ▸ Bottled Juices (BJC), ▸ Front-end-candies (FEC), ▸ Fabric Softeners (FSF),
- ▸ Laundry Detergents (LND), ▸ Refrigerated Juices (RFJ), ▸ Canned Tuna (TNA).

Törnqvist, 기하 Laspeyres, Laspeyres, Jevons 직접지수가 계산되고 기준기간은 이전연도, 12월 연결월로 매년 업데이트된다. 동적바스켓, 연쇄 Törnqvist 또한 계산되었고, 아래 표는 여러 방법의 평균지수, 그래프는 FSF(섬유유연제) 제품지수이다. 표와 그래프 다변지수는 공표지수에 연결 GEKS 25-HASP이고, 시뮬레이션을 통해 다음 사항을 얻을 수 있다.

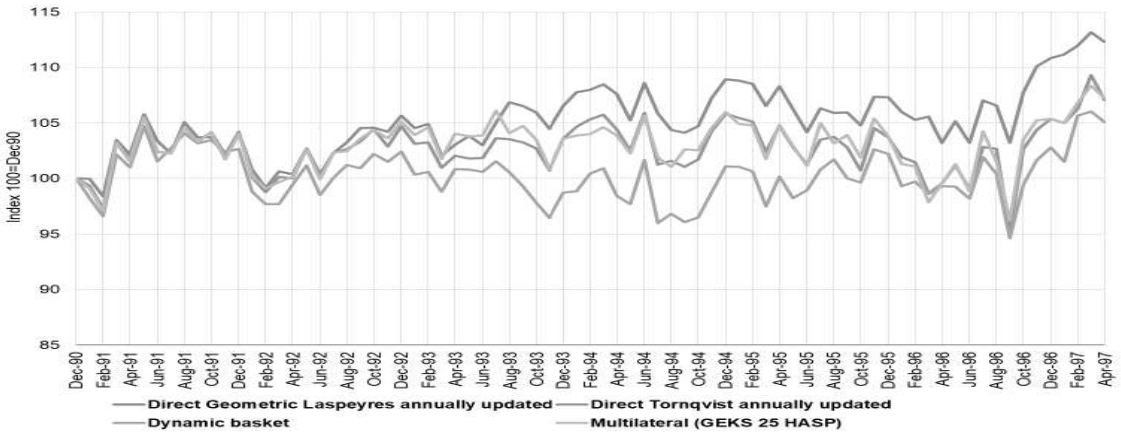
- ▸ 다변지수는 일반적으로 직접 Törnqvist 지수에 가깝다.
- ▸ 연쇄 Törnqvist는 직접 Törnqvist에 비교하여 하향 편향을 가진다.
- ▸ 직접 라스파이레스는 직접 기하 라스파이레스보다 높다.
- ▸ 동적바스켓이 다변지수와 유사하지만 약간의 차이점이 있을 수 있다.
- ▸ 직접 Jevons는 신뢰성이 낮아 보이며 다른 방법들(예: FEC, TNA 또는 RFJ)과 다를 수 있다.

표 : 평균지수(1990.12~1997.4), 1990.12=100

제품	연쇄 Törnqvist	동적 바스켓	직접 Törnqvist	GEKS 25 -HASP	직접 기하 Laspeyres	직접 Laspeyres	직접 Jevons
BJC	103.1	106.7	103.5	104.6	105.4	106.2	105.2
FEC	113.4	114.7	114.8	114.8	115.3	116	111.7
FSF	99.3	100.3	102.7	102.9	105.2	106.9	100.8
LND	89.2	94.1	94.9	96.8	92.3	94.9	96.8
RFJ	74.3	85.7	86.6	86.8	88.3	89.5	94.3
TNA	92.6	97.6	97.5	97.5	97.7	99.8	105.5

26) Index Compilation Techniques for Scanner Data, Claude Lamboray, UNECE ,Online meeting, June 2021

그래프 : FSF(섬유유연제) 지수(1990.12~1997.4), 1990.12=100



다변지수는 지수산식, window length, splicing 방법 조합으로 보여 줄 수 있다. 총 75개 방법이 테스트되었다.

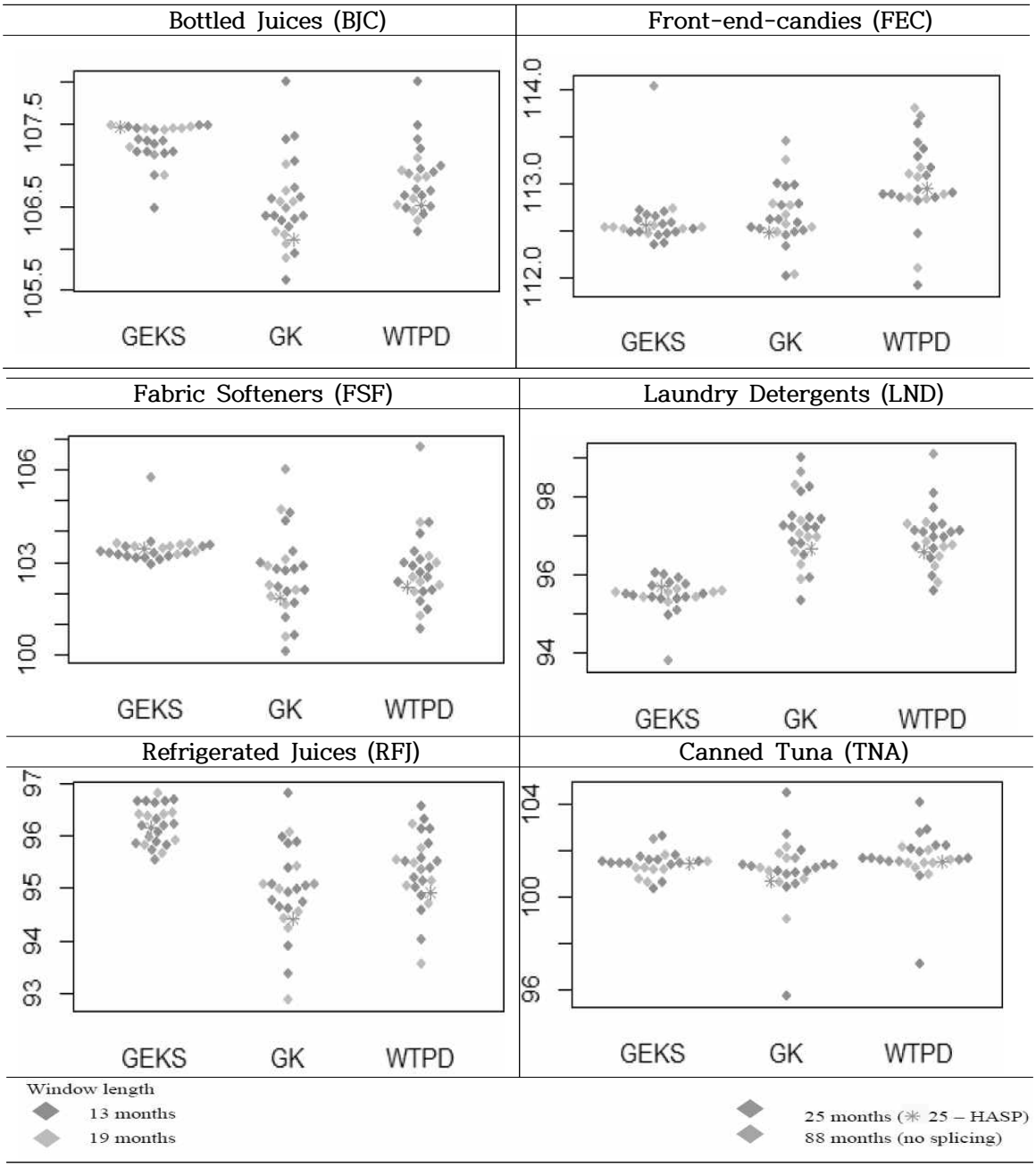
- 지수공식 3개: GEKS-Törnqvist, GK, WTPD
- 윈도우 길이 3개 (rolling time windows of 13 months, 19 months, 25 months)와 88 months (full time window)
- 스플라이싱 8개: movement splice, window splice, window splice on published indices, half splice, half splice on published indices, mean splice, mean splice on published indices, December splice

다음은 여러 다변지수 시뮬레이션 결과이고, 75개 평균지수(90.1~97.4월)는 그림에 표시하였고, 벤치마크는 GEKS 25-HASP[27] 지수이다.

- GEKS 진폭은 GK 및 WTPD보다 작은데, GEKS가 GK나 WTPD보다 윈도우 길이와 스플라이싱 선택에 덜 민감하다는 것을 보여준다.
- GEKS는 주어진 윈도우 길이 조건으로 여러 스플라이싱 방법 결과는 크게 다르지 않을 수 있다. GK와 WTPD의 경우, 주어진 윈도우 길이에서 여러 스플라이싱 방법의 결과가 더 크게 다를 수 있다.
- GK와 WTPD 패턴은 비슷해 보이지만 GEKS 패턴과는 다르다.
- 88개월 결과는 특이치로 보일 수 있다. 이는 세 지수공식 모두에 적용된다(예: FSF 참조). 이는 많은 연구에서 수행되는 전체기간에 걸쳐 산출된 다변지수가 적합한 벤치마크인지에 대한 의문을 제기한다.

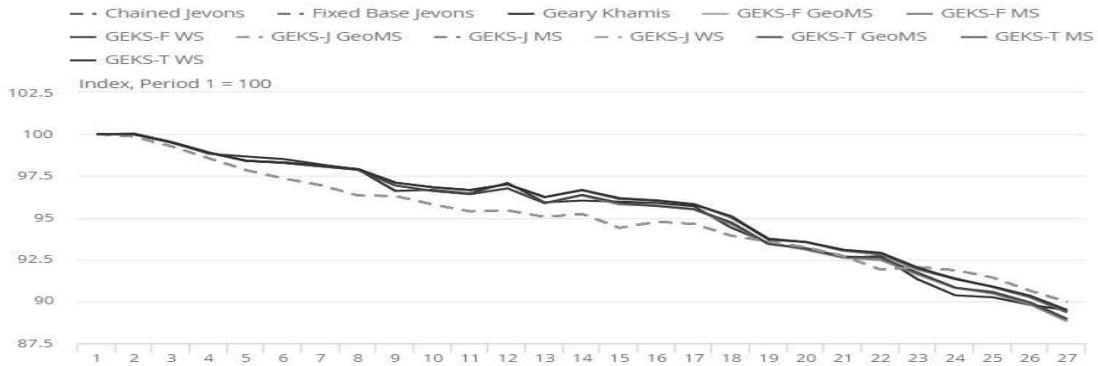
27) 25 month rolling time window, a half splice on published indices(25-HASP) GEKS

그림 : 다변지수, 평균지수 (1990.1~1997.4), 1990.1=100



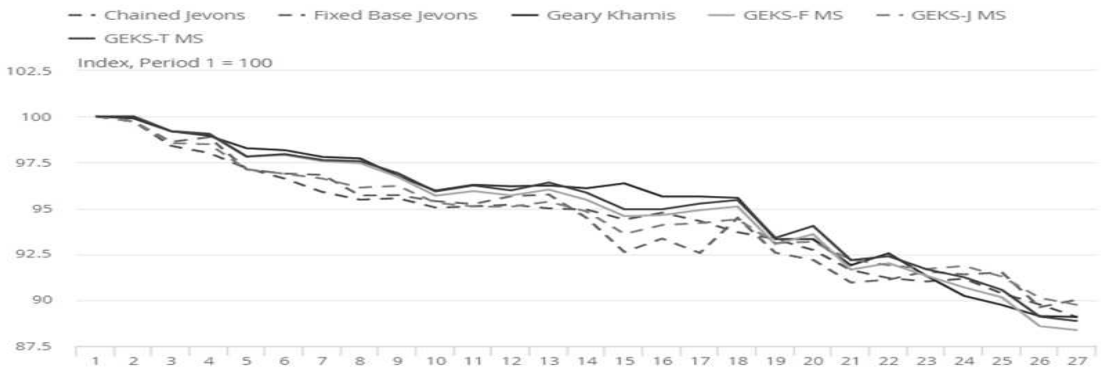
7.2. 영국 ONS[28)]

기본 자료(base synthetic data)



가중치 지수와 무가중치 지수 간의 차이가 가중치 방법 내에서 관측된 차이보다 평균적으로 더 크고, 제품 수준에서 가중치 존재여부가 가중치 지수 방법 자체의 선택보다 더 중요하다는 것을 보여준다.

높은 변동률(high attrition rate)

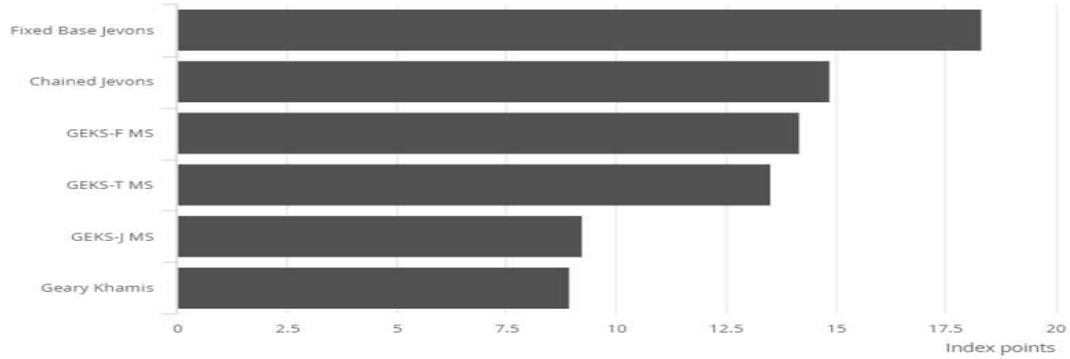


웹 스크래핑 및 스캐너 자료에서 주요 특징은 특히 의류시장이나 기술제품(PC, 노트북, 태블릿, 스마트폰 등)의 경우 시장에 진입하고 나가는 제품이 많다. 이 특징은 다변지수가 양변지수보다 더 낫다고 고려될 수 있다.

기본 데이터를 기반으로 한 지수 값과 높은 변동(churn)이 추가된 데이터 세트를 사용하여 산출한 지수 값 사이의 절대 차이 합계를 비교했다.

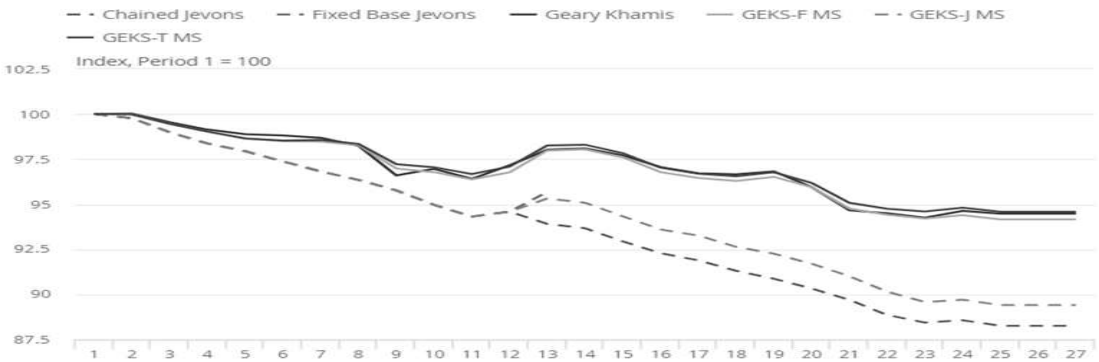
---

28) 출처 : 영국 ONS, New index number methods in consumer price statistics 1 September 2020



변동(churn) 높을 때 고정 및 연쇄 제본스는 다변지수보다 기본지수와 더 다르다. 고정 제본스는 매달 사용 가능한 데이터 비중이 계속 감소하므로 시간이 지남에 따라 연쇄 및 다변지수보다 품목을 덜 대표하게 된다. 또한 연쇄 양변지수는 일반적으로 연쇄편의로 인해 시간이 지남에 따라 기본 데이터 세트와 점점 더 다를 것으로 예상할 수 있다. 따라서 방법특성을 고려할 때, 시간이 지남에 따라 다변지수에 비해 고정기준과 연쇄 제본스에 대한 절대차이의 합이 증가할 것으로 예상할 수 있다.

제품 노후화(Product Obsolescence)



노후화는 제품이 기능적으로 다른 제품과 비교하여 경쟁에 밀려 소매업체가 그 제품의 생산이나 판매를 중단할 때 발생할 수 있다(예, 스마트폰). 고정 양변지수는 문제가 발생할 수 있는데, 기간 13이후 현재 및 기준기간에 더 이상 일치하는 제품이 없다. 따라서 더 이상의 지수 값을 생성할 수 없고 시계열은 종료된다. 가중치가 없는 방법은 가중치 방법보다 지수가 유의하게 낮아진다. 이것은 그들이 지출 정보를 사용하지 않기 때문이다.

Ⅲ. 웹스크래핑을 활용한 물가지수 작성[29)]

1. 웹스크래핑을 활용한 물가지수 작성

소비자들의 인터넷 구매와 웹사이트 운영 소매업체 수는 최근 극적으로 증가하였고, 웹은 거대한 정보 원천으로 성장했다. 우리는 일상생활의 많은 부분을 온라인에서 보내고 있기 때문에, 통계 생산자는 이 데이터를 간과해서는 안되고, 물가상승을 측정하기 위해 온라인 가격을 통합할 필요가 있고, 웹에서 가격을 수집하는 것은 비용이 많이 드는 대면 가격 조사와 통계단위의 부담을 줄여주기에 자원 관점에서 유리하다. 또한 고빈도로, 큰 볼륨으로 자료를 수집할 수 있기에 통계생산 프로세스를 가속화할 수 있다. 이런 관측치 증가는 통계 개선과 신 지표 개발에 큰 잠재력을 가지고 있고, 웹 스크래핑 기술을 사용하면 매일, 매시간 가격을 관찰할 수 있으며, 가격 변동에 대한 지식을 풍부하게 하고 수집 전략 애플리케이션을 향상시킬 수 있다. 가격 측정을 개선하는 메타데이터(예: 제품특성)도 수집할 수 있다. 인터넷 정보를 관리하기 위해서는 데이터를 보다 효율적으로 처리할 필요가 있다. 웹에서 HTML 형식으로 이용할 수 있는 정보를 중앙저장구조로 변환하여 분석과 조작을 보다 편리하게 하는 것이다. 이런 데이터를 웹에서 자동으로 검색하는 것을 웹 스크래핑이라고 한다. 이것은 보통 스크레이퍼라는 프로그램에 의해 서버에서 수행된다.

1.1. 웹 스크래핑 이점 및 한계

웹 스크래핑 이점은 다음과 같다.

- 데이터 수집 자동화로 수집 비용 절감, 가격 대표성 강화
- 가격 수집 빈도 증가(매일 vs 매월)를 통해 해당 기간 동안 더 대표적 가격을 제시, 평균 가격사용의 2차적 효과로는 변동성 감소
- 새로운 제품과 사라진 제품을 더 빨리 식별, 응답자 부담 감소
- 추출, 저장 및 명시적 품질 조정에 사용될 수 있는 풍부한 메타데이터 (즉, 제품 특성, 거래데이터와 같은 다른 데이터 소스도 보완가능)

29) (EUROSTAT 2021) How to start with web scraping in the HICP, ILO(2020)“Consumer Price Index Manual: Theory and Practice.”web scraping

웹 스크래핑의 지출정보 부족은 제품/제품 그룹들이 가중치를 부여 받을 수 없으며 사용할 수 있는 지수 공식이 제한된다. 웹 스크래핑의 기타 잠재적인 제한 사항은 다음과 같다.

- 온라인을 가진 소매점으로 제한(즉, 커버리지 부족 가능성)
- 웹 사이트가 변경되면 웹스크래핑이 실패할 수 있고, 기업이 웹스크래핑 활동을 탐지하고 이를 막기를 원한다면 인터넷 프로토콜 주소를 차단할 수 있는 등 이에 대비 등 정기적 IT 유지관리가 필요
- 통계청 내에서 수행되는 웹스크래핑은 정기적인 IT 유지 보수를 처리하기 위해 중간 프로그래밍 지식을 갖춘 컴파일러가 필요

1.2. 법률적 측면[30)]

웹 스크래핑의 법적 측면은 보통 국가마다 다르므로 국가별 적용규제를 검토할 필요가 있다. 다만 최근 중요한 대상인 만큼 기존 국내법 사례도 직접 연구해보는 것이 선호될 수 있다. 일반적으로, 유럽통계청은 유럽통계 목적을 위해 다양한 출처로부터 정보를 수집하고 접근할 수 있는 명확한 법적 권한을 갖는다.(유럽 통계 관행 강령 제2원칙)

2. 데이터 수집 및 데이터 액세스 권한

통계당국은 유럽의 통계목적을 위해 여러 데이터 출처로부터 정보를 수집하고 접근할 수 있는 명확한 법적 권한을 가진다. 행정, 기업, 가계 및 일반 국민은 법률에 의해 통계 당국 요청에 따라 유럽 통계 목적을 위한 데이터에 대한 접근 또는 전달을 허용하도록 강제될 수 있다.

2.1. 유럽 통계 개발, 생산 및 보급을 위하여 복수의 데이터 출처로부터 정보를 수집하고 접근할 수 있는 통계 당국의 권한은 법률에 명시되어 있다.

2.2. 통계당국은 법에 따라 행정자료를 신속하고 무료로 접근하고 통계목적에 사용할 수 있다. 통계목적에 보다 적합하도록 하기 위하여 행정기록 설계, 개발 및 중단에 처음부터 관여한다.

2.3. 통계당국은 법률행위에 기초하여 통계조사에 대한 응답을 강제할 수 있다.

2.4. 통계상의 비밀성과 데이터 보호를 보장하면서 비공개 자료와 같은 다른 자료에 대한 통계목적의 접근이 용이하게 된다.

30) Eurostat 법률적 측면

비록 이 권한이 통계 목적으로 웹사이트에 대한 접근을 용이하게 할 것이지만, 웹사이트 소유자들은 다른 이유로 스크래퍼의 데이터 접근을 차단할 수 있다. 그럼에도 상용 웹사이트들은 때때로 다른 특정 웹 스크래퍼들이 방문함으로 얻는 이점을 갖고 있기 때문에 서버들은 종종 이것을 고려하여 설계된다. 어떤 경우든, 가장 좋은 접근은 메타태그, 개인 정보 보호 또는 데이터베이스 보호와 관련된 웹 사이트 정책 관련하여 미리 소매점과 접촉하는 것이다. 좋은 관계를 구축하면 특히 스크레이퍼 에 영향을 줄 있는 웹사이트 변경의 경우에 도움이 될 수 있다. 통계청은 웹사이트 소유자가 가지고 있을 수 있는 정책뿐만 아니라 네티켓도 존중할 것이 권고된다. 네티켓을 따르기 위해서는 적어도 다음을 적용해야 한다.

- 스크레이퍼는 사용자-에이전트 문자열을 사용하여 식별됨
- 서버가 과부하 되지 않도록 요청 간에 충분한 일시 중지(예: 1초) 허용
- 스크래퍼는 야간 또는 한가한 시간에 운영[31)]
- 웹사이트가 로봇 제외 프로토콜을 사용하는 경우 존중되어야 함
- '자동 스크래핑 금지' 조항을 포함할 수 있는 웹 사이트 '약관' 섹션 존중

1.3. 웹 스크래핑 프로세스

통계청 웹스크래핑 수행 프로세스는 두 가지 주요 방법에 초점을 둘 수 있다.

① 통계 소프트웨어를 사용하여 통계청 내에서 수행되는 웹 스크래핑
② 제3자/사기업에서 얻어진 웹 스크래핑 서비스

통계청내에서 웹스크래핑을 수행하거나 외부업체 수행 선택은 내부 상황(예: 예산, 프로그래밍 용량, 유지보수 비용)에 따라 달라진다. 통계 소프트웨어를 통계청 내 수행 관련, 인터넷에서 정보를 스크래핑하는 프로세스는 다음과 같은 세 가지 단계로 나눌 수 있다.

① 웹사이트 스크래핑 허용여부 확인 ② 웹사이트 스크랩 ③ 수집 자료 정리

31) 야간 스크래핑의 단점은 스크래핑 과정에서 갑작스러운 버그를 해결할 수 없다는 것이고, 동시에 스크래퍼를 밤에만 가동하면 밤과 낮의 가격이 다를 경우 대표성이 운을 따를 수 있다.

웹사이트가 웹 스크래핑에 적합한지 여부를 결정하는 주요 방법이 있다. 웹 스크래퍼를 설치하는 직원은 웹사이트 이용약관 섹션에서 “사용조건“을 확인해야 한다. 여기서 웹 스크래핑이 허용되는지 금지되는지 여부를 지정하는 경우가 많다. 또한 robots.txt는 웹사이트 루트 디렉터리 내에 위치할 수 있고, 이 텍스트는 자세한 정보를 포함하고 있으며 웹스크래핑 허용된 사용자, 스크래핑에 사용할 수 있는 정보 및 금지된 것을 포함하여 웹스크래핑 조건을 개략적으로 설명한다. 웹사이트가 스크래핑이 가능하다고 결정되면 해당 사이트에 스크레이퍼가 설정되고, 웹사이트의 카테고리 구조를 찾아 스크래핑할 관련 카테고리를 식별한다. 프로그래머는 포함되고 제외될 구조 부분들을 정의한다. 예를 들어, 웹사이트는 모든 개별제품 범주를 나열한 다음 그 개별범주의 제품을 복제하여 “신제품(new products)” 등과 같은 추가 범주를 나열할 수 있다. 이러한 추가 범주는 프로그래머가 제외할 수 있다. 그 다음 스크래퍼는 그 인터넷에서 모든 제품과 가격을 다운로드한다.

정보를 가져오는 데 사용할 수 있는 두 가지 옵션이 있다. 첫째, 데이터는 웹 사이트에서 복사하여 붙여넣기 위해 스크레퍼를 사용하여 텍스트로 수집될 수 있다. 이 경우 스크래핑 후 텍스트 클리닝하는 절차가 수행된다. 대안으로 HTML 태그와 클래스를 사용하여 제품과 가격을 끌어낼 수 있으며, 이는 데이터를 추출하고 정리하는 데 있어 보다 표적화된 접근방식이다. 이 방식의 다른 이점은 제품식별자가 HTML에서 숨겨질 수 있으므로 태그를 사용하여 가져와서 제품설명에 추가할 수 있다(텍스트 설명에 의존하는 것과 달리). 그러나 HTML 태그의 사용이 모든 웹사이트에서 쉽지 않다. 텍스트로 수집된 정보의 경우, 제품집합과 가격만 남겨지도록 자료가 정리될 필요가 있다. 제품과 가격을 “noise“에서 분리하기 위해 데이터패턴을 파악하고 코딩해야 한다. 가격지수를 구성하는 데 이용 가능한 정보와 관련하여, 표1.3.은 통계청에 의해 스크랩된 일반적인 메타데이터 요약이고, 데이터 프레임에는 일반적으로 다음이 포함된다.

- 일자 : 스크랩 특정일(날짜)
- 유통업체명 : 소매업체명(text)
- 카테고리 : 소매업체 웹사이트 분류(text)
- 제품 ID : 제품설명(text)
- 가격 : 제품의 특정가격(숫자)

표 1.3. 웹스크래핑, 일반적인 자료 구조[32)]

Date	Retailer	Category	Product ID	Price
July 10, 2019	Retailer ABC	Children's Shirts	Brand XYZ– Short Sleeve Polo Shirt	$45.00
July 10, 2019	Retailer ABC	Children's Shirts	Brand XYZ– S/S Regular Shirt	$55.00
July 10, 2019	Retailer ABC	Children's Shirts	Brand XY– Short Sleeve Regular T-Shirt	$15.00
July 10, 2019	Retailer ABC	Children's Shirts	Brand XYZ– Long Sleeve Regular Shirt	$65.00
July 10, 2019	Retailer ABC	Children's Shirts	Brand XYZ– Short Sleeve Regular Shirt	$35.00

1.4. 품목 적용 범위 및 표본

1.4.1. 품목 적용 범위

웹 스크래핑할 COICOP 하위지수 선택시, 고려해야 할 여러 사항이 있다. 이는 비용 절감과 관련이 있을 뿐만 아니라 방법론 이유(예: 품질 조정이 거의 필요하지 않은 품목; 온라인 구매가 점점 더 대표적인 품목; 가중치 정보가 스크랩될 수준에서 이용 가능한 품목)도 포함될 수 있다. 대부분 품목에서 스캐너 데이터가 간단히 더 선호되는 데이터 소스이다.

통계청이 웹 스크래핑 프로젝트를 시작해야 하는 경우, 가장 좋은 전략은 주요 잠재적 어려움이 해결 될 때까지 스크래핑을 점진적으로 도입하는 것이다. 프로세스가 조직화되고 지속 가능하게 숙달되면 더 많은 품목으로 확장 할 수 있을 것이다. 이런 관점에서 현재 인력을 병행 교육하거나 IT 기술을 갖춘 전문가, IT 기술 업무 경험이 있는 졸업생을 고용하는 것 또한 나중 단계에서 성과를 거둘 수 있다. 스크래핑이 유용할 수 있는 분야는 연료, 전기, 보험, 서적, 의류, 의약품, 전자제품, 항공요금, 숙박이고. 이들 중 몇 가지는 제품의 높은 다양성과 특성으로 많은 분류 문제를 일으킬 수 있으며 캡차를 동반하는 경우가 많기 때문에 최상의 시작은 아닐 것이다. 관측치 수가 많을수록 필요한 통계 방법(대체, 품질조정, 그룹핑)에도 영향을 미칠 수 있다.

32) ILO(2020). "Consumer Price Index Manual: Theory and Practice." web scraping

1.4.2. 표본

적합한 웹사이트를 선택하기 위해서는 중요한 요소들을 고려해야 하는데, 스크래핑 결정전에 다음과 같은 웹사이트 특성을 평가해야 한다.

대표성 : 인구, 소비 관점에서 대표범위는 중요한 차원이고 온라인 수집 가격이 실제 전자상거래를 대표한다면, 조사원이 아울렛을 방문할 필요가 없도록 구분해야 할 것이다. 제품 관련, 웹사이트와 실물상점에 있는 것이 동일하지 않거나 가격이 다를 수 있다는 것을 주의해야 한다.

볼륨(volume) : 관측치 수와 사용 가능 변수들, 특히 이번이 첫 스크래핑 프로젝트인 경우, '많을수록 좋다'는 것이 모든 경우에 적용되는 것은 아니다. 이후에 관측된 가격수를 변경하면 최종 결과가 어떻게 달라질 수 있는지 확인하는 것도 필요할 수 있다.

콘텐츠 원본 : 모든 사이트가 원본 콘텐츠를 제공하는 것은 아니고, 종종 사이트는 다른 사이트로부터 수집된 자료를 제공한다. 실제 데이터 소유자와 스크레이퍼 사이에 이런 추가계층이 있으면 문제가 발생할 수 있다.

Site 안정성: 스크랩 자료의 지속성은 웹 사이트 안정성에 크게 좌우되며, 정기적 변경 대상이 아니며 웹에서 사라질 가능성이 없는 사이트를 스크래핑하는 것이 좋다.

기술적 특성: 사진이나 pdf 가격은 스크랩하기에 어려울 수 있으며, 정적 사이트는 동적 콘텐츠가 있는 사이트에 비해 시작하기가 더 쉽다. 메뉴 구조가 명확한 사이트를 선택하면 스크래핑도 쉬워진다.

방법론 고려사항 : 웹소스를 통계 단위별로 직접 쿼리할 수 있는 것이 좋고, 데이터를 사용 가능한 품목에 연결하면 분류에 도움이 될 수 있다.

대상 변수 : 변수식별은 항상 사전에 보장되는 것이 더 좋으며, 선택된 변수가 이용 가능한 다른 웹소스와 결합될 수 있는지도 미리 생각해보는 것이 좋은데, 이는 새로운 지표 개발까지 이어질 수 있기 때문이다.

메타데이터 : 최적 계층화 및 균질성을 만들려면 자세한 제품설명과 함께 메타데이터 제공 사이트를 스크래핑하는 것이 좋다. 가격과 메타 자료 수집은 다른 일정을 가진 다른 스크래퍼로서 더 잘 작동할 수 있다.

보다 일반적인 권장 사항으로, 적합한 사이트를 찾았으면 API에 액세스할 수 있는지 또는 안정적으로 역할을 할 수 있는 다른 사이트가 있는지 확인할 필요가 있다. 내려야 할 결정 중 하나는 스크래핑 빈도이다. 인터넷에서 수집된 가격은 시간이 지남에 따라 매우 변동적일 수 있다. 예상되는 가격 변동성이 더 많은 추출을 의미할 수 있지만, 대부분의 경우 일일 또는 그 빈도가 더 낮은 기준으로 추출하기에 충분해야 한다. 스크래핑이 안정적일 때까지 빈도를 조정할 수 있는 것이 좋고, 고장(자동 재시작 스크립트) 또는 불완전한 데이터 수집(부족한 값을 impute 옵션)의 경우에 대응할 시간이 있어야 할 것이다.

1.5. 분류 및 데이터 검증

1.5.1. 분류[33)]

대량 웹 스크랩 자료를 효율적으로 처리하려면 제품 자동분류가 가장 중요한 과제일 것이다. 제품(product-offers)은 균질제품으로 그룹핑되서, 지수 계층화 일부인 제품범주에 할당되어야 한다. 이는 장기적으로 관리 유지 되어야 할 분류규칙이고, 이 규칙설계는 일반적으로 수동으로 처리되고, 분류를 위한 일반적 전략은 텍스트문자열 검색을 수행하는 것이다. 스크랩된 제품설명에서 특정 키워드가 추출된다는 의미다. 웹사이트가 사용하는 상품 카테고리와 CPI 상품 카테고리를 매핑할 수도 있다. 데이터 세트가 작으면 전문가가 제품을 수동으로 분류할 수도 있다. 이런 점에서 더 현명한 규칙이 필요하다. 훈련(training) 자료 세트의 그 알려진 제품 분류를 고려하여 제품을 분류하는 감독 학습 알고리즘, 자동분류기법이 가격 수집 비용을 더욱 크게 줄일 수 있을 것으로 기대된다.

---

33) (ILO 메뉴얼) 웹스크래핑 자료는 통상 계층분류(예: COICOP)에 매핑되는데 필요한 일부 기본제품 텍스트 및 범주설명을 포함하고 있어, 분류를 해결하기 위해 고려해야 할 접근방식은 다음과 같다.

- 텍스트 문자열 검색 : 분류를 위해 설명 문자열에 키워드가 있는지 여부를 확인
- 카테고리 매핑 : 데이터 세트(또는 그 일부)에는 각 제품에 대한 소매업체 범주들을 포함. 이러한 범주들 중 하나가 CPI 분류에 포함되는 경우 그 범주는 그 분류에 매핑될 수 있음
- 수동 매핑 : 설명 문자열을 살펴봄. 이 옵션은 소규모 데이터 세트에 대해 가장 실현 가능한 옵션
- 지도학습 알고리즘: 자동분류를 위해 텍스트와 훈련결정 사이 패턴을 식별하는 통계학습 알고리즘에 훈련 데이터를 제공

다음은 스크랩된 데이터에서 추출하여 분류를 위해 집합을 만들 수 있는 일반적인 파라미터의 예이다.

- ID – 품목 고유주소(예: URL) 또는 코드
- 제품유형 – 분류(예: 남성, 여성, 아동복)
- 이름 – 제품이름이 반드시 고유할 필요는 없음
- 설명 – 품목별 특성 ▪ 가격 – 사이트에 표시된 '오퍼 가격'

동일한 제품이 시간에 따라 추적되도록 통계청은 GTIN 또는 상점들이 사용하고 있는 다른 품목코드를 따를 수 있다. ID는 시간에 따라 고유하고 안정적인 것이 중요하다. 다른 품목인데, 동일한 ID가 있는지 확인하는 것도 좋은 방법이고 맞춤법 오류, 약어 또는 용어뿐만 아니라 다른 언어 버전을 검사하는 것도 수행될 수 있다. 짧은 품목 서술계층과 어휘 질이 좋은 잘 구성된 웹 사이트를 선택했다면, 분류가 더 쉬울 것이다. 그럼에도 분류를 개선하거나 유용한 상세한 품목설명을 추가하기 위한 방법을 연구하는 것이 좋다. 보다 상세한 분류가 일반적으로 물가지수의 편성을 개선한다.

1.5.2. 모니터링 및 데이터 검증

선택도구의 타당성과 유용성, 스크래핑 적용 메커니즘을 계속 모니터링해야 하는데, 이는 무언가 잘못되었을 경우 다시 데이터를 수집하기 위해 그 시점으로 돌아가는 것이 종종 불가능하기 때문이다. 자료수집 과정을 자동으로 모니터링하여 중요한 차이가 있을 경우 경고가 발생하도록 하는 것이 합리적이고, 모니터링 대시보드를 추가하면 스크래퍼가 제대로 작동하는지 확인할 수 있다. 문제를 탐지하는 좋은 proxy는 데이터 크기 또는 분류되지 않은 품목수가 될 수 있다. 비가용성 사례 처리나 서버 문제 알림 제공, 중단 기간에 대비해 사전에 백업 솔루션을 준비하는 것이 좋다. 필요한 경우 스크래핑을 다시 시작하거나 다시 할 시간을 주기 위해 모니터링이 조기에 이루어져야 한다. 또한 사이트의 유지 관리된 부분에서 가격을 수집하는 것이 좋다. 일부 문제들은 그 스크랩된 데이터에 수천 개의 관측치가 포함되어 있는 경우 특히 발견하기 쉽지 않다. 따라서 중복 레코드, 의심스러운 값, 이상치, 0과

같은 값을 갖는 품목 수 등을 모니터링 하는 등 데이터 수집 중에 표준 검사를 도입하는 것이 좋다. 데이터 검증은 물론 정가와 세일가격 차이, 계절성 등 품목 관련 특성을 고려하는 것도 좋다. 처음에는 기존 가격 수집과 어느 정도 병행하여 비교하는 것이 좋다.

1.6. 지수 산출 및 자료 통합

지수 산출을 위한 입력, 각 스크랩된 가격에 대해 다음과 같은 추가정보를 이용할 수 있다고 가정한다 ; 스크랩된 날짜/시간, 스크랩된 웹사이트 이름, 제품식별자, 제품설명 및 제품에 대한 기타 메타 정보(예: 제품범주)
각 데이터는 제품(product-offer)에 해당하는데, 이는 웹사이트에서 특정 시간에 가격을 관측할 수 있는 특정 제품이다. 따라서 제품은 관측된 가격이 집계되어야 하는 다음과 같은 차원으로 구성된다.

- 시간 차원(스크래핑 날짜/시간)
- 대상처 차원(웹 사이트)
- 제품 차원(제품 ID, 제품 설명, 제품 범주)

일반적으로 지수 산출은 다음의 세 가지 단계에 따라 수행된다.

① 개별제품(individual product) 정의에 관해 선택을 해야 한다. 제품은 시간, 웹사이트 및 제품에 걸쳐 집계 될 필요가 있다.
② 기초물가(elementary price)지수는 개별제품을 종합하여 구한다. 지수 공식은 가중치, 제품특성을 사용할 수 있고 사용하지 않을 수도 있다.
③ 기초물가지수는 하위분류 수준(예:COICOP)에서 정의된 지수구조 내에서 집계된다.

1) 1단계: 개별 제품(individual product) 정의

먼저 시간 경과에 따른 집계를 분석한 다음 제품 간 집계를 분석한다. 이 두 단계가 모두 필요하지 않을 경우 개별제품(individual product)은 제품(product-offer)과 일치하며 그 다음 단계로 진행할 수 있다.

시간별 집계(aggregation across time)

데이터는 한 달에 한 번 이상 스크랩되고, 한 달 동안 가격 변동이 심한 것으로 알려지면 고빈도(예: 매일)로 수집하는 것이 좋다. 만약 한 달 동안 모든 시점이 소비자에게 동일하다면, 전체 기간은 균일한 것으로 간주될 수 있다. 개념적으로 단가가 원하는 목표가격이 될 것이다. 그러나 각 제품에 대해 가중치를 사용할 수 있는 것은 아니다. 따라서 실제평균가격 계산은 불가능하다. 여기서 주어진 달의 스크래핑된 가격들의 비가중 산술평균 또는 기하평균 중 하나를 가지고 산출한다. $p_{i',t'}$는 제품(product-offer) 관측 가격을 나타내며, 여기서 i'는 특정 상품, t'는 특정 시점, T'는 한 달 동안의 스크랩된 가격수이다.

$$p_{i',t} = \frac{1}{T'}\sum_{t'} p_{i',t'} \text{ 또는 } \quad p_{i',t} = \prod_{t'}(p_{i',t'})^{\frac{1}{T'}}$$

$p_{i',t}$는 '시간에 걸쳐 집계된 제품가격'이고. 이 가격은 웹 스크래핑 시기와 빈도에 따라 달라지며, 각 제품이 동일한 횟수만큼 판매된다고 가정한다. 기준 월을 대표하는 평균가격을 내놓는 것이 목표다. 매월 동일한 시점에 가격을 수집하는 것이 좋다.

제품 간 집계(aggregation across products)

일반적으로 제품 ID는 자료에서 쉽게 사용할 수 있는 가장 세분화된 제품수준이며 비슷한 것과 비슷한 것을 비교하는 것을 확실히 하기 위해, 동일한 제품 ID를 가진 제품(시간 집계된) 가격을 추적할 수 있다. 이는 유효한 전략이 될 수 있으며, 이 경우 제품 간에 집계할 필요가 없다. 상세한 제품 특성이 이용될 수 있고 헤도닉 모형(2단계 참조)에서 사용될 경우에는 제품 간 통합이 필요하지 않다. 그러나 제품 ID는 각 개별 제품들의 품질 변동 없이 변경될 수 있다. 이런 경우 기존 ID와 변경 ID 사이의 가격 변동은 포착되지 않는다. 할인가격에 상품 아이디가 체계적으로 시장에서 사라지면 물가지수가 하향 편중될 위험이 있다. 따라서 데이터에 높은 변동(attrition rate)이 있고/또는 모델 매칭 접근법이 편향된 결과로 이어진다면, 시간 집계된 제품들을 그룹화하고 더 광범위한

동질제품을 구성하는 것이 권장될 수 있다. 이런 전략은 특히 의류와 신발에 적합할 수 있다. 동질제품 구성은 대개 데이터 중심이며 환경(예: 제품유형, 제품 특성 가용성 등)에 좌우되며 실제로 다음 단계들이 수행 될 수 있다.

① 웹 사이트에서 스크랩된 메타데이터(예: 웹 사이트별 제품범주)뿐만 아니라 시간 집계 제품을 설명하는 텍스트를 탐색적으로 분석한다.
② 변수 목록과 그룹화에 사용될 수 있는 양식을 도출한다. 예를 들어, 시간 집계된 제품은 브랜드, 제품유형 또는 기타 특정 제품특성에 따라 그룹화 될 수 있다. 또한 가격 자체를 차별적 변수(예: 고가, 저가 제품)로 사용할 수도 있다[34].
③ 각 시간 집계된 제품에 대해 스크랩된 데이터에서 정보를 추출하여 이러한 보조 변수를 구성한다.
④ 일정한 값이나 양식을 취하는 변수를 적용하여 균질한 제품을 정의한다. 예를 들어, 동질 제품은 브랜드=X 및 제품유형=Y인 시간 집계 제품으로서 정의 될 수 있다. 좀 더 엄격한 정의는 브랜드와 제품 유형뿐만 아니라 일부 다른 제품 특성에도 제한을 두는 것이다.

목표는 대략 동일 품질의 시간 집계 제품으로 구성되는 동시에 시간 내내 이용 가능한 동질적인 제품을 설계하는 것이다. 시간에 따른 안정성과 균질성 사이에서 절충이 이루어져야 한다. 균질제품가격은 주어진 달의 균질제품의 정의에 해당하는 시간에 걸쳐 집계된 제품 가격 $p_{i',t}$의 산술평균 또는 기하평균으로 구해진다.

$$p_{i,t} = \frac{1}{N_t}\Sigma_{i'}\, p_{i',t} \text{ 또는 } \quad p_{i,t} = \Pi_{i'}\left(p_{i',t}\right)^{\frac{1}{N_t}}$$

균질 제품의 목표가격은 단가이다. 시간 집계 제품 집합이 시간에 따라 다르기에 이것을 '비매칭'이라고 부른다. 주어진 달의 누락(비매칭) 가격들이 같은 달 관측된 가격들의 산술(또는 기하)평균과 동일하다고 가정될 경우 표준 듀토 또는 제본스 지수가 된다.

34) 가격 수준을 사용하여 품질을 측정하는 것은 특히 가격 변화가 주요 측정 대상이기 때문에 신중하게 수행되어야 한다. 가격이 동일하다고 해서 반드시 동일한 품질을 의미하는 것은 아니다.

앞에서 설명한 개별 제품 정의(Eurostat 2021)와 비교해 다음과 같이 ILO 웹스크래핑 개별 제품(Individual Products) 정의를 살펴보고자 한다.

가격측정의 필수 부분은 품질변화와 신제품 도입을 고려하는 것이다. 제품품질이 변하면, 지수가 순수한 가격변화를 반영하도록 조정된다는 것은 웹스크래핑에서 또한 중요하다. CPI 표본의 제품 출현과 소멸은 품질에 상응하는 어떤 변화를 적절히 처리되지 않는 한 지수를 편향시킬 가능성이 있다. 이는 데이터에 포함된 많은 가격, 높은 제품 변동(attrition), 제품의 수명주기 시작과 종료 시점의 비정상적 가격 움직임을 보이는 경향으로 인해 모든(또는 대부분의) 웹스크래핑 가격을 통합하는 지수 계산에 문제를 제기한다.

이 문제를 다루는 한 가지 접근은 양 시점 제품만 사용하여 두 기간 사이의 가격변동을 추정하는 것이며, 따라서 신제품과 소멸제품의 가격을 배제하는 것이다. 소멸상품은 남은 재고를 할인 가격으로 판매되는 경우가 있으며, 비슷한 품질의 상품과 연계되지 않을 경우 "재출시" 문제가 발생할 수 있으며 지수 하락편향이 발생할 가능성이 있다. 이 문제는 첨단 기술 상품, 의류 등 여러 제품에 대한 연구에서 확인되었다.
이 문제를 극복하기 위해, 주로 텍스트에서 특성 정보(예: 브랜드와 셔츠 유형)를 추출하여 광범위한 제품정의를 형성하는 데 초점을 맞춘 몇 가지 실용적 전략이 제안되었다. 제안된 주요 기법은 다음과 같다.

- 광범위한 제품 범주의 사용(예: 표1.3의 아동 셔츠)
- 텍스트/정규 표현 함수: 반구조적 텍스트에서 특성을 추출하는 데 사용(예: 표 1.3.의 “XYZ“는 특성 “브랜드“를 추출).
- 근사(fuzzy) 일치 함수: 페널티 함수를 사용하여 텍스트 문자열을 근사적으로 일치시키는 데 사용
- 김독학습 알고리즘 : 제품 분류를 위해 텍스트와 훈련결정 사이의 패턴을 식별하는 통계학습 알고리듬에 훈련 데이터를 제공
- 감독되지 않은 학습 알고리즘 : 특징(예: 텍스트 문자열 및 가격)과 알고리즘을 사용하여 제품 “클러스터“를 자동으로 정의

2) 2단계: 기초물가지수(elementary price indices) 산출

개별제품이 구성되고, 가격이 계산되었다면, 그 개별 제품에 대한 집계를 수행한다. 개별제품이 하나의 범주로 분류되었다고 가정하면, 목적은 이 범주의 기초물가지수를 산출하는 것이고 여러 옵션들이 있다. 웹 스크랩 자료의 지수 산출 한계는 가중치가 없다는 것이다. 개별 제품들이 얼마나 자주 판매됐는지는 대개 알 수 없기 때문에 물가지수가 편향되지 않은 방식으로 생산될 수 있도록 주의가 필요하다.

가중치 없는 집계

가중치는 일반적으로 웹스크랩 자료에서는 사용할 수 없어, 가장 표준 접근은 기준기간(예: t-1의 12월)과 비교시점 이용 가능한 개별제품 가격을 비교하는 고정기준 제본스를 사용하는 것이다.

$$I_{t,0} = \prod_{i \in N} \left(\frac{p_{i,t}}{p_{i,0}} \right)^{\frac{1}{|N|}}$$

이 접근에서는 기준월 또는 비교월에 이용할 수 없는 개별제품을 고려하지 않는다. 1년이 지나 기준기간 업데이트시 전년에 등장한 개별 신상품이 포함될 수 있다. 보다 동적인 접근은 매월 기준 기간을 업데이트하고 매월 링크를 통해 연속적인 지수를 얻는 것이다. 그러나 이 접근은 예를 들어 할인 가격 경우 등 편의를 초래할 수 있으므로 주의해야 한다. 다른 옵션은 하향 편의를 줄이기 위해 GEKS-Jevons 등 다변지수를 사용하는 것이다. 가중치 없는 지수를 구현할 때는 웹 스크랩 가격이 얼마나 대표적인지 이해하는 것이 중요하다. 대형 데이터세트에는 매우 인기 있는 모델과 거의 판매되지 않는 모델이 모두 포함될 수 있다. 대표할 수 있는 가격 변동을 얻기 위해서는 추가 정제 작업 및 표본 추출이 필요할 수 있다.

가중치를 사용한 집계

개별제품 가중치는 알려져 있지 않을 지라도 대략 추정될 수도 있다. 한 가지 방법은 개별제품의 기초가 되는 제품(product-offers)이 스크랩된 횟수를 가중치 대안으로 사용하는 것이다. 따라서 지출 유형 가중치는 주어진 개별 제품 i에 대해 다음과 같이 도출될 수 있다.

$$w_i = \Sigma_{t'} \Sigma_{t'} p_{i',t'}$$

웹사이트 인기순위 활용 등 가중치를 도출하는 기술도 있다. 개별제품 집계는 이런 가중치를 사용, 라스파이레스 또는 기하라스파이레스를 적용할 수 있다.

$$I_{t,0} = \Sigma_{i \in N} \frac{w_i p_{i,t}}{\sum_j w_j p_{i,0}} \text{ 또는 } \quad I_{t,0} = \Pi_{i \in N} \left(\frac{p_{i,t}}{p_{i,0}} \right)^{\frac{w_i}{\Sigma_j wj}}$$

보다 발전된 접근은 각 개별제품에 월별 가중치를 할당하고 다변지수를 적용하는 것이다. 가중 지수는 비가중 지수보다 선호된다. 그러나 프록시 가중치적용 기초지수의 신뢰성은 면밀히 모니터링 되어야 한다.

제품특성(characteristics)을 가지고 집계

특정 제품(예: 전자제품)의 경우 개별제품의 특성과 가격을 함께 스크랩하는 것이 종종 가능하다. 이를 통해 헤도닉지수를 산출 할 수 있고, 웹 스크랩 데이터 맥락에서, 시간 더미 헤도닉을 사용하는 것이고, 제품 특성 외에 시간더미가 설명변수로 추가되는 회귀분석에 기초한다.

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t$$

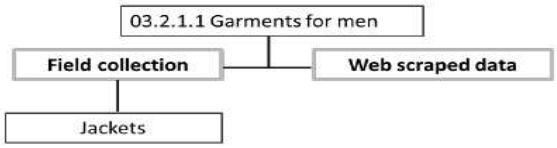
여기서 z_{ik}는 개별제품 i의 특성 k에 해당하며, D_i^t는 그 개별제품이 t 기간에 사용 가능한 경우 1이고 그렇지 않은 경우 0인 더미 변수이다. 공선성[35]을 피하기 위해 한 달 동안 더미 변수가 없다(예: 기준 기간).

35) 2개 이상의 설명변수 간에 하나가 완전한 선형관계가 존재하면 이를 공선성이라 하며, 하나 이상의 완벽한 선형관계가 있을 경우, 이를 다중 공선성(multicollinearity)이라고 한다. 회귀분석에서는 설명변수간에 선형성이 존재하면, 특히 공선성에 근접하면 통계적 추론이 불안정하게 된다. 만약 두 변수간에 높은 선형성이 존재하면, 각 변수들의 독립적으로 종속변수에 미치는 영향을 분리하기 매우 어렵게 된다.

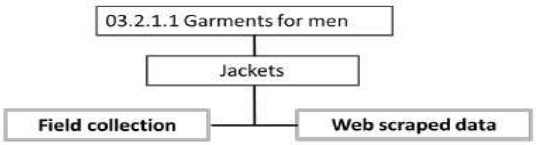
기준기간에 대한 비교기간 t의 지수는 다음과 같이 구한다. $I_{t,0} = \exp(\hat{\delta}^t)$ 개별제품 세트는 기간마다 다를 수 있고 일반적으로는 가중치가 없고, 모델이 OLS(ordinary least squares)를 사용하여 추정된다. 이 접근은 제품(product-offers)을 균질 제품으로 그룹화할 필요없이 대신 제품 특성이 모델에서 직접 사용된다.

3) 3단계: 분류(예: ECOICOP, COICOP) 아래의 집계 구조

웹 스크랩 데이터가 소비자물가지수에 통합되어 다른 데이터 소스와 결합되는 수준결정이 내려져야 한다. 예를 들어, 이 데이터가 COICOP(또는 ECOCOP) 5자리 하위 클래스 바로 아래에 통합될 수도 있고 대안으로, basket에 이미 포함된 더 자세한 제품 범주에 첨부될 수도 있다.



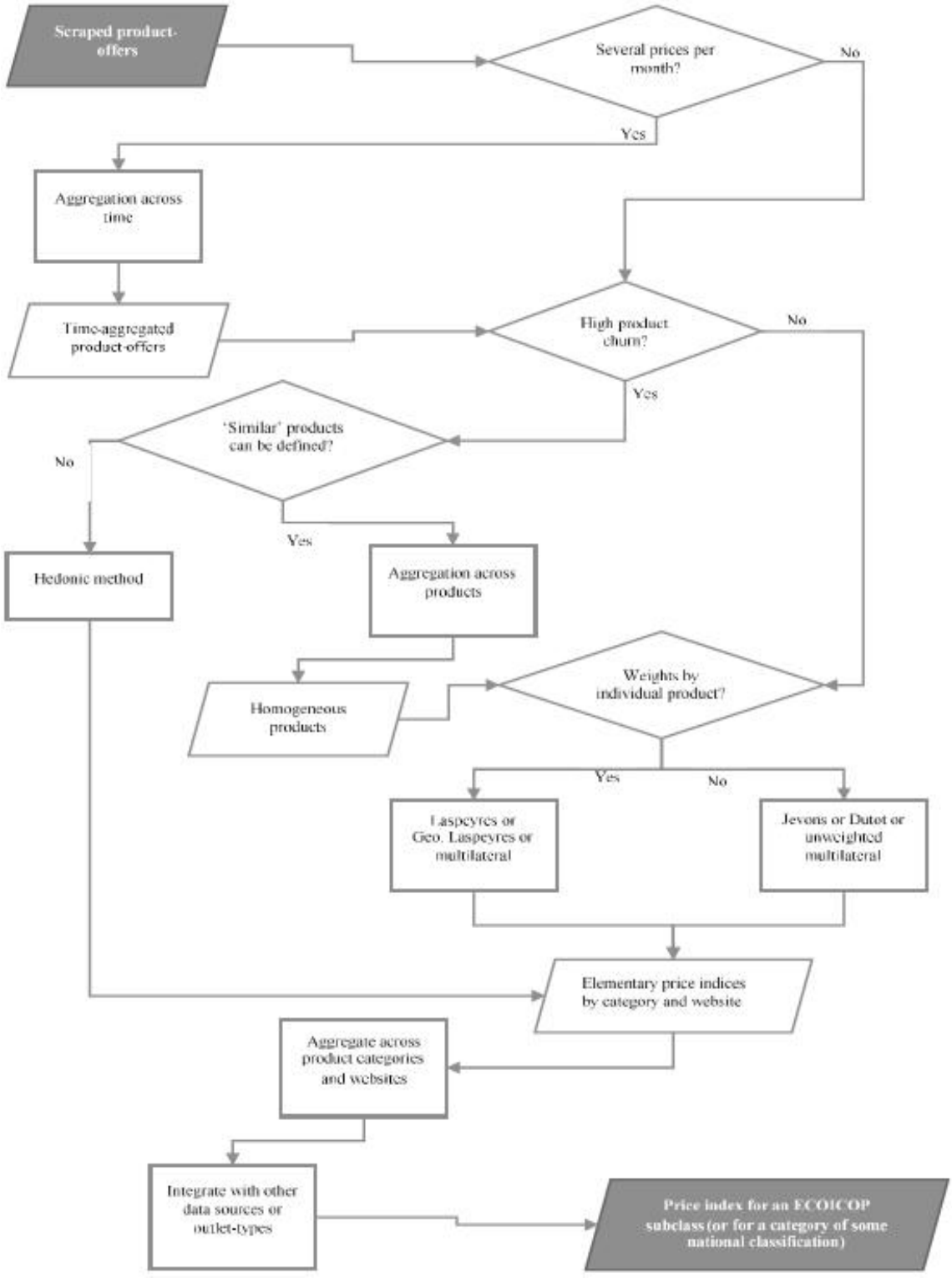
Integration at ECOICOP5 level



Integration at a product category

더 높은 수준의 통합은 더 많은 유연성과 웹 스크랩 데이터의 이점(예: 더 넓은 제품 적용 범위) 고려할 수 있다. 낮은 수준으로 통합하면 예를 들어 현장 대면 수집한 가격과의 일관성을 더 높일 수 있다. 여기서, 웹 스크랩 데이터는 가중치가 할당되어져야 하는 계층에 해당된다. 여기 가중치에 대한 데이터 소스는 가계 예산 조사 또는 소매업 통계가 될 수 있다. 예를 들어, 웹에서 구매한 것과 현장 아울렛에서 구매한 것은 구별이 될 수 있다.

다음 플로우차트[36]는 지금까지 살펴본 웹스크래핑 지수 산출과정을 나타낸 것이다.



36) (EUROSTAT 2021) How to start with web scraping in the HICP

2. 캐나다 통계청 CPI 웹스크래핑 활용 사례

2.1. 의류 및 신발 지수 : 웹 스크래핑 데이터 통합[37)]

CPI에 적용되는 데이터 출처와 방법을 주기적으로 검토, 업데이트하고 있다. '20.1월 의류 및 신발 구성 요소에 웹스크래핑 데이터를 통합하였다. 의류 및 신발은 '17년 CPI 바스켓의 5.17%를 차지하며, 의류, 신발, 의류 액세서리, 시계 및 보석 등 4가지 구성요소 지표로 구성되고, 이들 지수의 일부 가격을 선별된 소매업체에서 웹 스크래핑 하고 있으며 더 이상 현장에서 수집되지 않는다. 이런 변화로, 통계청은 매달 의류 및 신발 소매점에서 약 10% 더 적은 가격을 조사한다. 의류와 신발 부문 현장 가격은 2주 동안 전국에서 조사원들이 한 달에 한 번 기록한다. 웹스크래핑은 한 달 내내 가격을 수집할 수 있고, 현장에서 수집한 데이터와 결합되어 보다 강력한 데이터를 만들 수 있다. 표본소매업체는 동일하지만, 웹에서의 가격 정보 접근성 및 웹 스크래핑 효율성은 더 다양한 제품에 대한 인구사 수준의 가격 데이터를 수집할 수 있게 하고 제품 적용 범위를 개선한 것이다.

2.2. 항공운임 지수 개선[38)]

항공운임 지수는 캐나다 내 도시와 국제선 목적지 사이의 항공운임 평균 변화를 추정한다. '17년 가중치 기준 1.49%를 차지한다. CPI 데이터 소스에 대한 지속적인 검토로, 항공운임 지수의 품질을 개선하기 위한 2단계 이니셔티브가 완료되었다. 2018년 3월 도입한 1단계는 항공운임 지수에 온라인으로 항공여행을 쇼핑하고 구매하는 것을 포함하도록 하여 다음과 같이 개선하였다.

- 항공사 및 도시로 증가
- 여행기간 표준화
- 사전 항공료 예약기간을 나타내는 예약시차 도입

37) The Integration of Web-Scraped Data into the Clothing and Footwear Component of the Consumer Price Index(Statistics Canada, February 19, 2020)
38) Enhancements to the Air Transportation Index in the Consumer Price Index (Statistics Canada, January 22, 2020)

20년 1월 도입한 2단계는 인터넷 가격 데이터가 GDS[39]의 API를 이용해 자동으로 수집되는 가격으로 대체하면서 항공사와 목적지가 더 늘어났다. API는 항공예약 시스템을 GDS와 통합하여 여행사와 소비자가 온라인으로 티켓을 검색하고 예약할 수 있도록 한다. 출발지, 도착지, 출발일, 귀국일 등의 입력 요청 매개변수를 수집하고 기준 요금, 세금, 여행일정 세부정보 등의 데이터가 포함된 응답파일을 제공한다. 통계청은 항공운임 지수 산출에 사용되는 데이터를 얻기 위해 여행업에서 가장 큰 두 개의 GDS에 가입, 이 데이터는 API를 통해 광범위한 항공요금과 여행일정정보에 대한 액세스를 제공한다. 항공료 수집 API사용에 기초한 가정은, 여기서 얻은 자료는 소비자들이 이용할 수 있는 항공편과 가격을 대표한다. API는 이용 가능한 최저요금(이코노미석)을 반영하고 있다. 표는 API에서 발생하는 적용범위, 수집빈도 및 가격수 요약이다.

표 1 API 데이터 1 단계 및 2 단계 도입 결과

	1 단계 (인터넷 가격 수집)	2 단계 (API 데이터)
도시 pair의 수	53	180
예약 시차 횟수	2	4
매월 수집 빈도	3	매일
컬렉션 당 가격 (도시 pair 당)	2	최대 15
항공사 수	4	캐나다 시장에 서비스를 제공하는 모든 항공사
매월 사용된 가격	636	API 당 300,00 이상

2단계에서, API수집으로 항공운임 지수샘플은 캐나다에 서비스를 제공하는 모든 항공사를 포함하도록 확장되었다. 가격은 각 도시조합에서 중요한 항공사를 대표하도록 공항활동조사 매출액을 사용하였고, 도시조합 수가 180개로 증가하였다. 각 도시 pair, 예약시차 및 수집 일에 대해, 중요한 항공사 중 하나 이상의 운임을 포함하여 15개의 가장 저렴한 운임관측치를 선택한다. 선정된 180개 도시 pair 가격 지수는 국내, 아시아, 태평양 등 5개 여행지 분야를 대표하는 기본 집계로 분류된다.

39) GDS(Global Distribution Systems : 항공예약 판매시스템)는 여행서비스 제공업체와 소비자 간 거래를 활성화하고 촉진하는 기업이다. 이들은 온라인 여행사의 주요 데이터 출처 및 예약 컨택 포인트를 나타낸다. GDS는 광범위한 서비스 제공업체로부터 여행데이터를 수집 및 통합하고 여행사가 비행기 좌석 예약, 렌터카 대여, 호텔 예약 등을 하도록 한다. 따라서 여행서비스를 준비할 때 수백 개의 항공사, 호텔 및 기타 최종 제공 업체와 직접 연결할 필요가 없다.

항공운임 지수는 일일 API요청을 통해 얻은 가격을 사용하여 계산되고 포괄범위와 수집빈도 증가는 지수를 계산하는 데 사용되는 가격수의 현저한 증가를 가져왔다. 결론적으로, 항공료 데이터 수집을 위한 API로의 이전은 항공사와 여행지의 샘플을 대폭 확대하고, 가격 수집빈도를 높이는 계기가 되었다. 이러한 변화는 CPI 항공운임 지수의 품질과 관련성을 유지하기 위한 중요한 개선이다.

Ⅳ 캐나다 및 미국 소비자물가지수 작성 방법

1. 캐나다 소비자물가지수

1.1. 작성 방법[40)]

CPI는 시간 경과에 따른 고정 바스켓 비용을 비교하여 소비자 물가변동을 나타내는 지표이며, 바스켓에는 동등한 수량 및 품질의 상품 및 서비스가 포함되기에 순수한 가격변동을 반영한다고 정의하고 있다. 대상인구에는 모든 도시 및 농촌 가구에 거주하는 개인이 포함되고 집단가구와 인디언 보호구역에 거주하는 사람, 외국과 그 가족을 대표하는 공무원, 화이트호스, 옐로나이프, 이칼루트 이외의 유콘, 노스웨스트 준주, 누나부트 거주자는 제외한다. 지리적 범위는 모든 10개 주, 화이트호스, 옐로나이프, 이칼루트가 해당된다. CPI 상품과 서비스 품목적용 범위에 소득세, 자선기부금, 연금기부금, 저축과 투자, 공공지원 건강보험 프로그램으로 제공되는 생명보험과 건강서비스는 해당되지 않는다. 분류체계는 CPI부서에서 가계지출조사 및 「Office of Privacy Management and Information Coordination」와 협력, 표준상품분류에 기초하여 개발되었다. 가격유형은 조사시점에 소비자가 부담하는 가격을 반영(예외: 사용자비용 기준 자가주거비)하고, 조건 없는 보조금과 할인혜택이 반영되며, 사례별로 리베이트를 고려한다. 조사시기는 품목특성에 따라 많은 자료가 월별로 가격이 조사되지만, 일부 자료들은 월별로 수집을 하지 않고, 조사는 해당 월 첫 3주에 걸쳐 진행한다.

40) IMF DSBB https://dsbb.imf.org/ , Statistics Canada CPI Detailed information

가중치 관련해서는 기본등급, 품목범주 가중치는 가계최종소비지출(HFCE)에 기초하고 하위수준 가중치는 다른출처(스캐너 데이터 등)에서 얻을 수 있다. 품목은 기본등급의 가격변동을 대표하고, 적정기간 사용할 수 있어야 하며, 세부규격은 지속적으로 검토되며 언제든 변경 될 수 있다. 대상처 관련 임대료는 계층화된 랜덤 표본추출, 다른 품목의 경우 점포유형, 판매량 및 도시내 위치 등을 고려하여 판단 표본추출을 적용한다. 표본크기는 매달 약 10,000개의 임대료 자료와 최대 7,300개의 대상처에서 110,000여개의 가격자료를 수집한다. 임대료는 가구에서 조사하고 다른 가격 수집은 방문, 행정자료(스캐너 자료 포함)와 인터넷을 사용하며, 일부 가격은 전단이나 전화를 통해 수집한다. 세부규격에는 제품특성이 자세히 설명되어 있고, 상당히 좁은 품질범위를 지정 가능하고, 경우에 따라서 특정모델이 식별된다.

자료처리 관련, 누락된 가격은 전월 가격이 이월되거나 같은 최저수준 상품분류에 속하는 다른 상품의 평균가격 변동을 사용하여 대체된다. 상품전문가는 소매업체와 제조업체로부터 받은 정보, 제품 및 시장 지식을 바탕으로 품질차이를 추정하고 Hedonic은 컴퓨터 장비에 사용되며 의류 및 기타 제품에 대해 연구 중이다. 교체(replacement) 관련 조사자는 규격에 맞는 다른 아이템을 찾도록 요청받고, 새 상품은 규격에 부합해야 하며, 이전 조사된 것과 비슷해야 한다. 신규 조사상품(item)은 신제품이 충분한 시장 점유율을 획득할 때 그 제품 규격이 가격샘플에 추가되고, 계절 품목은 비수기 상품에 대한 집단추정(group imputation)이 특히 의류에 사용된다. 자가주거비는 사용자비용 접근법이고, 구성요소는 주택담보대출 이자비용, 주택소유자 대체비용, 재산세, 주택소유보험, 주택소유자 유지 및 수리비, 기타 소유 주거비이다.

계절조정 자료는 다음과 같이 13개 계열이 전국수준에서 생산되고 있다. • 전체 품목 CPI • 8가지 주요분류 • 4가지 특수집계: ① 중앙은행 정의 변동성 높은 요소 8개와 간접세 변동 효과 제외한 모든 품목, ② 중앙은행 정의 변동성이 높은 요소 8개 제외한 모든 품목, ③ 식품 제외 모든 품목, ④ 식품 및 에너지 제외 모든 품목

현재 생산되는 계절조정은 각 계열이 직접 조정되는 것으로 계절조정된 하위 구성요소를 집계한 결과가 아니다. 계절조정지수는 수정 가능하고, 매월 전월 계절조정지수가 수정 대상이다. 지난 3년간 계절적으로 조정된 값은 매년 1월분 공표시 보정된다

기준년도는 2002=100이고, 가격자료 연결은 이전 달과 매칭하며, 필요시 품질조정을 하며. 최소 단위지수는 평균가격(비가중 기하평균 또는 일부 산술평균) 변화를 이전 지수에 연결하여 계산한다. Lowe 산식은 기본지수를 상위레벨로 집계하는 데 고정 가중치를 적용하며, 가격 변화에 대한 가중치 업데이트 방법으로 산출된 지출 가중치는 연결월 가격을 사용하여 갱신한다. 조사자는 가격 변동을 확인하여 비정상적 움직임에 대한 설명을 제공하고 가격이 합리적인지 본부에서 확인한다. 외부 정보를 가지고 상품 전문가 및 경제학자가 계산을 검토하고, 내부 물가지수 품질보증팀과 외부 물가측정자문위원회가 있어 CPI 모든 측면을 검토한다. 보도시기는 매월, 익월 16-22일 근무일 중이며, 8개 주요 분류, 180개 기본 상품 지수, 그리고 21개 특별분류 지수로 세분되며, 이 계열의 4분의 3은 지방자료별로도 수록된다. 이는 홈페이지 온라인 게시판 또는 데이터, 영어와 불어로 제공된다. 향후 15개월간의 공표일은 매년 12월에 발표하며, 또한 최소 4개월간의 공표일은 통계청 인터넷 웹사이트에 게시한다. 매월 보도 공표는 모든 이해관계자가 이용할 수 있도록, 당일 08:30 EST 캐나다 통계청 인터넷 웹사이트에 게시한다.

1.2. 소비자물가지수 자주 찾는 질문(FAQ)[41]

CPI는 생계비 지수(COLI)인가? CPI는 생계비에 근접하기 위해 종종 사용되어 왔으나 CPI와 COLI는 직접적으로 비교될 수 없다(동일하지 않다). CPI는 전국 가구의 평균 소비 습관을 나타내는 상품과 서비스의 고정된 바구니에 기초하여 소비자가 직면하는 소매가격의 평균 변화를 측정한다. COLI는 일정한 생활수준을 유지하는데 있어 소비자가 경험하는 가격 변화를 측정하는 것이다.

41) Statistics Canada Consumer Price Index: Frequently asked questions

데이터는 수정되는가? CPI는 물가연동 목적으로 광범위하게 사용되기 때문에 개정될 수 없다. CPI 정확성은 만약 수정 가능한 CPI를 도입한다면, 수정치는 때로는 위로, 때로는 아래로 나타날 수 있다. 결과적으로, 그런 개정이 있을 때마다, 공공 및 민간 부문은 과소 또는 과다 지불하도록 요구될 것이다. 개정으로 인한 불확실성으로 인해 임금과 계약상 지급이 발생할 때 최종 지급으로 간주할 수 없다면 일반적으로 경제에서 비용이 더 높아지게 될 것이다.

모든 주와 지역 가중치를 동등하게 부여하고 있는가? 각 주와 지역의 물가 변동은 캐나다 전체 소비자 지출에 대한 해당 주의 상대적 중요도에 따라 가중치를 부여한다. 예를 들어 온타리오는 40.59%의 지출 점유율을 가지고 있는데, 이는 캐나다 전체 가계소비지출의 40.59%를 차지한다는 것을 의미한다(2020년 basket weights 기준).

계절조정 자료는 무엇인가? CPI에 어떤 영향을 미치나? 계절조정은 원지수에서 계절적인 가격 변동을 분리한 다음 이를 제거한다(일반적으로 매년 거의 동일한 규모로 발생하는 계절적 및 달력 영향)는 것이고, 이는 경제의 "기조적인" 소비자 물가 인플레이션을 얻기 위해서이다.

직접 계산을 위해 CPI raw 데이터에 액세스할 수 있는가? CPI 계산을 위해 수집된 정보는 통계법에 따라 기밀사항이다. 따라서 통계청은 입수 정보를 그 개인, 기업 또는 단체의 사전에 서면 동의 없이 누설하는 것을 법으로 금지하고 있다. 따라서 원시데이터에 대한 액세스를 제공할 수 없다.

다른 기준년도를 갖도록 지수를 변환하려면 어떻게 해야 합니까? 지수를 다른 기준기간으로 변경하려면 변환계수를 계산해야 하고, 이 계수를 계산하려면 기준기간으로 지정할 연도의 연평균 지수가 필요하다. 예로서, 기준기간 1986=100을 사용하여 2009년 3월 전체 CPI를 구하려면 현재 기준기간(2002=100) 1986년도의 연평균 지수 값은 65.6이다. 변환계수는 연평균 지수를 100으로 나누는데, 65.6/100 = 0.656이므로 변환

계수는 0.656이다. 계산하고자 하는 월의 지수를 변환계수로 나눈다.

114.0(2009년 3월 기준기간 지수 2002=100) / 0.656(변환계수) = 173.8

2009년 3월 CPI는 173.8(1986=100)이며, 변환계수를 곱하기만 하면 원래 지수(2002=100)로 돌아간다. 173.8(1986=100) x 0.656 = 114.0(2002=100)

1.3. Basket 개편 및 연쇄지수[42)]

바스켓 가중치는 주로 특정 기준연도의 가계최종소비지출에서 도출된, 실제로는 혼합지출 가중치이며, 이는 지출의 가격과 수량이 다른 기간에서 발생한다는 것을 의미한다. 이 혼합지출 가중치는 CPI 고정 바스켓 개념에 필수적이다. 가구지출조사(SHS) 추정치를 CPI 제품 및 지리적 분류와 일치시켜 기초집계에 대한 가중치를 도출한다. 그러나 SHS가 상세정보를 제공하지 않는 경우가 있으므로 바스켓 가중치는 일부 경우에 대체소스로 구성된다. 주택소유자 대체비용과 주택담보대출 이자비용 기초지수 가중치는 보완 자료를 필요로 하며, 또한, 기타 통계청 조사, 행정자료 및 소매업 스캐너 데이터를 포함하는 대체소스를 사용하여 SHS가 세부사항을 제공하지 않는 제품등급에 대한 지출을 추가로 세분화한다. 술과 담배에 대한 지출은 SHS추정치가 소매판매 및 정부 소비세수자료보다 낮기 때문에, 이는 SHS에서 과소 보고되는 경향이 있어 보완 데이터로 편의로 의심되는 특정 SHS지출 추정치를 수정한다. 바스켓 갱신시, 가중치로 사용된 지출을 평가하기 위해 Bortkiewicz-Szulc decomposition 방법을 사용해서, 지출 가중치의 신뢰성을 평가하기 위해 수량 및 가격의 상대적 변화를 비교한다. 지출자료의 품질평가는 CPI 기본분류[43)] 수를 결정하는 데에도 도움이 된다. 기본분류(basic classes)는 소비지출 데이터의 가용성과 품질뿐만 아니라 기초집계(elementary aggregates) 내 지출 분포 안정성에 기초하여 결정된다. 예를 들어, 기초집계 내에서 소비지출의 분포가 자주 변화하는 경우, 그 지출에 대한 새로운 정보를 이용할 수 있을 때 지출 가중치 수량을 갱신하는 것이 유리할 수 있다. 이 경우, 통계청은 기본분류가 바스켓 수명 동안 수량

42) The Canadian Consumer Price Index Reference Paper(Statistics Canada, 2/27/2019)
43) basket 기간 동안 수량 가중치를 고정하는 제품과 지리적 분류 교차점에서 가장 낮은 수준

이 갱신될 수 있는 기초집계보다 높도록 지정한다. 바스켓 업데이트 간에 기본등급 수준 이하의 수량을 변경하는 관행은 소비지출에 대한 새로운 정보를 적시에 통합할 수 있다는 점에서 편익을 제공한다.

바스켓을 갱신하는 과정은 기초집계에 할당된 가중치를 현행 소비지출 패턴을 대표하는 것으로 만드는 것이다. 과거에는 CPI바구니가 가장 최근 SHS의 새로운 지출 데이터를 사용하여 4-5년마다 업데이트되었다. 2011년 바스켓 업데이트를 시작으로 가중치는 2년마다 갱신된다. 가중치 업데이트 및 품질확보 외에도 바스켓 업데이트는 다음을 포함한 지수의 다른 측면을 검토하고 업데이트할 수 있는 기회를 제공한다.

- 제품 및/또는 지리적 분류를 보다 대표적으로 변경
- 대표 제품 및 대상처 샘플 검토 및 업데이트
- 기초집계(elementary aggregate) 수준 이하 가중치 갱신
- 기초지수(elementary indices)에 대한 방법과 개념 검토.
- 배포용 문서 및 제품 업데이트

CPI는 고정 바스켓 지수를 연쇄하여 계산된다. 즉, 일련의 고정 바스켓 지수가 연속적 시계열을 생성하기 위해 연결되었음을 의미한다. 이는 바스켓 갱신시 중단되는 것을 방지하기 위해 필요하다. 바스켓 간 연쇄 지수는 바스켓 갱신 시점에 이루어진다. 바스켓 간 지수를 연쇄적으로 연결하기 위해서는 신구 바스켓의 혼합지출 가중치를 공통기간의 가격(링크월, 연결월)으로 표현한다. 이 연결월 가중치는 연결월 가격으로 표현되는 혼합지출을 얻기 위해 원래 지출 가중치를 가격 갱신하여 구한다. CPI 바스켓 기준기간 b는 1년이므로, 연결 월의 월간 혼합지출을 얻기 위해서는 가중치 조정이라는 프로세스가 필요하다. 링크 월의 월간 혼합 지출은 두 단계로 계산된다.

첫째, 바스켓 기준연도 b의 연간지출은 바스켓 기준연도 평균 가격으로 나눈다. 두 번째 단계에서는 바스켓의 고정 수량 값을 링크 달의 가격으로 표현하기 위해 초기 값을 링크 월로 갱신한다. 새 바스켓에 대한 링크 월 혼합 지출이 확보되면 이를 사용하여 지수를 계산한다.

[Lowe 산식]

$$P_{Lo}(p^0, p^t, q^b) \equiv \frac{\sum_{i=1}^{n} p_i^t q_i^b}{\sum_{i=1}^{n} p_i^0 q_i^b} = \frac{\sum_{i=1}^{n} (p_i^t / p_i^0) p_i^0 q_i^b}{\sum_{i=1}^{n} p_i^0 q_i^b} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) s_i^{0b}$$

$$s_j^{0b} = \frac{p_i^0 q_i^b}{\sum_{j=1}^{n} p_j^0 q_j^b} = \frac{p_i^0 q_i^b (p_j^0 / p_j^b)}{\sum_{j=1}^{n} [p_j^b q_j^b (p_j^0 / p_j^b)]}$$

바스켓 연계 월 다음 달에 새로운 바스켓을 사용하여 계산된 물가지수에 이전 바스켓에 대해 발표된 지수 수준을 곱한다. 연쇄 연결은 분류별로 수행되고[44], 현재 CPI는 2002=100 지수이고, 2002년 CPI는 1996년 바스켓에 기초했다. 이후 바스켓 갱신이 10번 있었다.

・2002.12월 링크된 2001년 바스켓, ・2004.6월 링크된 2001년 개정 바스켓,
・2007.4월 링크된 2005년 바스켓, ・2011.4월 링크된 2009년 바스켓,
・2013.1월 링크된 2011년 바스켓, ・2014.12월 링크된 2013년 바스켓,
・2016.12월 링크된 2015년 바스켓, ・2018.12월 링크된 2017년 바스켓,
・2021.5월 링크된 2020년 바스켓, ・2022.4월 링크된 2021년 바스켓

예를 들어, 2017 바스켓의 도입에 따라, 2002=100의 지수 기준기간을 갖는 모든 연쇄지수는 9개의 고정 바스켓으로 이루어진 연쇄이다.

$$I_{chained}^{2002,t} = I_{2017}^{Dec2018,t} \times I_{2015}^{Dec2016,Dec2018} \times I_{2013}^{Dec2014,Dec2016} \times I_{2011}^{Jan2013,Dec2014} \times I_{2009}^{Apr2011,Jan2013} \times I_{2005}^{Apr2007,Apr2011} \times I_{2001r}^{Jan2004,Apr2007} \times I_{2001}^{Dec2002,Jun2004} \times I_{1996}^{2002,Dec2002}$$

지수 12개월 변동이 두 바스켓에 걸쳐 있는 경우, 즉 두 비교기간(기간 t와 기간 t-12) 사이에 바스켓 갱신이 수행된 경우 변동 기여도의 계산은 다음과 같다. 이는 바스켓에 걸쳐 연쇄지수가 둘 이상의 고정 바스켓을 사용하여 계산되기 때문이다. 두 바구니에 걸쳐 연쇄로 연결된 물가지수의 12개월 변동 기여도는 두 부분으로 계산한다. 첫 번째는 구 바구니에 관한 것이고 두 번째는 신 바구니에 관한 것입니다. 바스켓 전체의 기여도를 도출하기 위해 연환지수를 사용해야 한다. 두 개의 바스켓에

44) 바스켓에 걸쳐 지수를 연쇄적으로 연결하는 방법은 지수가 각각의 하위 지수의 직접 평균이 되지 않는다. 이로 인해 지수의 수준이 하위지수의 범위를 약간 벗어나게 될 수 있다.ILO (2004)

걸쳐 있는 지수의 12개월 변동분 기여도를 계산할 때, 구 바스켓 기여도의 합계와 신 바스켓 기여도의 합계가 반대 기호(+/-)를 가질 수 있다.

$$\left(\frac{I_A^{0,t}}{I_A^{0,t-12}}-1\right)=\left[\sum_i\left(\frac{I_i^{0,link}}{I_i^{0,t-12}}-1\right)\times w_i^{t-12,old}\right]+\left[\sum_i\left(\frac{I_i^{link,t}}{I_i^{link,link}}-1\right)\times w_i^{link,\neq w}\times I_A^{t-12,link}\right]$$

$I_i^{link,link}=100$ 여기서, $w_i^{t-12,old}$: t-12 시점에서 품목(i) 구가중치

$w_i^{link\_\neq w}$: 연결월(중첩월)에서 품목(i) 새로운 가중치

$I_A^{t-12,link}$: 기준기간 t-12 대비 링크월에서 지수

1.4. 2020년 기준 CPI 가중치 개편[45)]

최근에는 2021년 가중치 개편과 2022년 가중치 개편이 있었고, 2021년 가중치 개편(2021.7.21.)을 중심으로 설명하고자 한다.

COVID-19은 지출방식에 큰 영향을 미쳤으며, 이러한 구매 패턴변화를 포착하는 것은 캐나다인이 경험하는 가격변동을 CPI가 반영하도록 보장하는 데 중요하고, 통계청은 CPI를 포함한 핵심 통계가 정확하고 시기적절하며 최고 품질을 갖도록 하기 위해 그 방법을 갱신하고 있다.
Laspeyres 유형의 Lowe 물가지수로서 수량은 상위수준 집계를 위해 가중치 기준기간에 고정된다. 주어진 집계의 바스켓 비중이 클수록, 그 집계의 가격 변동은 전 품목 CPI에 더 많은 영향을 미칠 것이다.
지금까지 가중치는 주로 가계지출조사(SHS)[46)]에서 도출되었다. 2021.6월 기준 가중치를 발표(21.7.)하면서 CPI에 사용된 가중치와 분류 구조는 '17년 SHS를 대체하여 '20년 가계최종소비지출(HFCE) 소비자 지출 패턴에 기초하여 업데이트 되었다. '20년 대부분 특징은 COVID-19가 가계지출에 큰 영향을 미쳤다. 원유가격이 하락하기 시작했고, 항공여행과 육로 통행 제한으로 관광분야에 큰 혼란이 있었다. 바이러스 확산을 제

45) An Analysis of the 2021 Consumer Price Index Basket Update, Based on 2020 Expenditures, Statistics Canada, July 21, 2021

46) CPI basket 구성비는 품목 분류 차이로 인해 총지출의 SHS 구성비와 다를 수 있음. 또한 CPI의 적용범위에 포함되는 일부 소비 비용은 SHS에 포함되지 않음. SHS외에도 국민 계정 데이터, 소매 판매 데이터 및 인구 조사 데이터와 같은 다른 지출 출처 자료가 CPI basket 가중치를 계산에 할용

한하기 위해 물리적 거리 대책이 제정되었고 식료품을 더 많이 사고 휘발유와 옷을 덜 사게 되었다. 레크리에이션, 관중 관람 엔터테인먼트 등 '21년에 거의 소비가 불가능한 일부 상품과 서비스에 대한 지출은 0에 가깝게 떨어졌다. 대유행으로 인한 예상치 못한 소비습관 변화로 인해 당초 계획한 2019년 가계지출에만 의존하는 것이 아닌 보다 최근의 2020년 지출을 포함하기 위해 가중치 갱신이 '21.6월로 연기되었다.

전국 가계최종소비지출(HFCE)에 기반한 첫 바스켓 업데이트이고, 2020년 HFCE외에 가장 최근의 2019년 SHS와 지방 HFCE도 데이터 품질을 개선하고 기본 방법을 강화하기 위하여 CPI 바구니 가중치에 통합되었다. 이 자료는 지역에 대한 고품질 추정치를 제공하며, SHS의 상세 데이터 및 2020년 다른 대체 데이터를 사용하였다. 국민계정(SNA)의 HFCE를 CPI 지출 가중치의 주요 소스로 통합하기로 한 결정은 COVID-19이전에 이루어졌으며 개념, 일관성 및 실용성에 대해 수행된 광범위한 연구와 분석의 결과이다. SHS는 다른 가계조사와 마찬가지로 측정과 표본추출에 어려움을 겪을 수 있기 때문에, SNA 자료 통합은 소매판매와 같은 다른 경제지표와의 일관성을 더한다. 또한 적용범위에서 SNA와 SHS 사이에는 중요한 차이가 있다.

< 표1 적용 범위 비교: 가계지출 및 국민계정(SNA) 조사 >

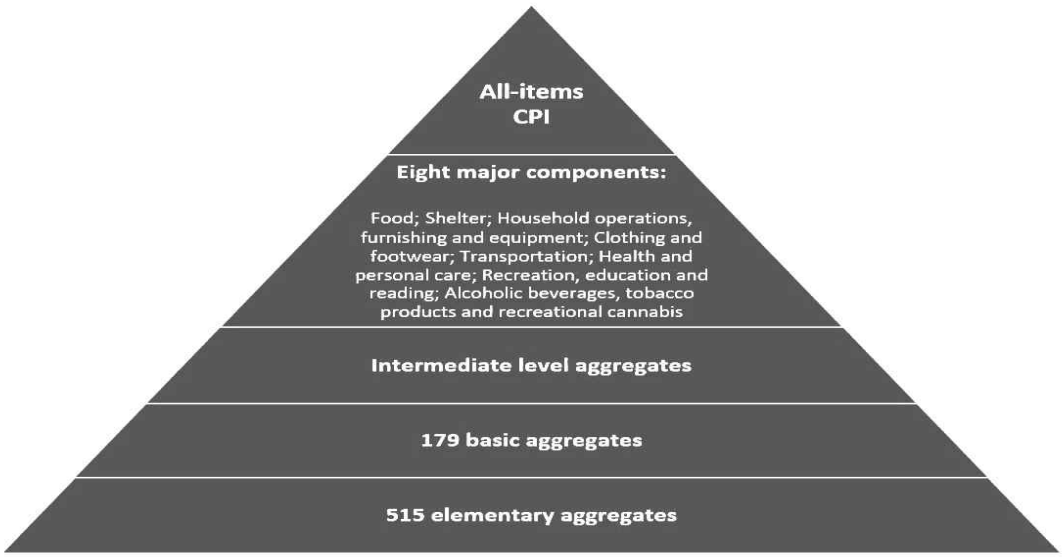
	가계지출조사(SHS)	국민계정(SNA)	품질 보증
목표 집단	국내 소비 측정	캐나다 내(가구, 집단 가구 및 비거주자) 국내 지출 측정	해외소비, 거주지 외 소비, 외국인 지출 제외를 위해 거주 조정 필요[47]
제품 분류	고유한 분류 구조	목적에 따른 개인소비 분류 기준(COICOP)	전문가 검토 후 제품분류일치 수행
대상 지역	퀘벡, 온타리오, BC 하위 주 세부 지역을 포함 19개 지역	캐나다의 모든 13개 주 및 준주를 포함	SHS 하위 주 세부 정보가 사용
개정 정책	수정 불가	SNA cycle 내 수정 가능	체계적 수정을 제거하기 위해 평가 수행 각 SNA 공개에 포함된 엄격함을 통해 전체 품질이 보장될 수 있음[48]

47) A residential adjustment is required to exclude consumption abroad, consumption outside of the province of residence and expenditures by foreigners

48) Overall data quality can be guaranteed through the rigour embedded in each SNA release.

SHS(2019년)와 SNA(지역 계정 HFCE, 2019년) 자료가 2019년 소비지출에 대한 종합적 개요를 제공하고, 2020년 바스켓은 2020년 분기별 HFCE 국가 수준 데이터를 모두 통합하였다.

상품 및 서비스의 소비자 물가 지수 분류는 5단계 하향식 계층 구조에 따라 구성된다(아래 도표 참조).



상단에는 8개의 주요 구성요소[49]가 포함된 전품목 CPI, 8개 주요 구성요소 아래에는 자가주거비, 차량운영과 같은 중간 수준의 집계가 있으며, 기본 집계는 179개이고, 일반적으로 하나 이상의 기초집계를 취합한 결과이며, 515개 기초 집계의 많은 부문은 미발표된다.

시간이 지남에 따른 소비패턴이 변경됨에 따라 바스켓에서 기초집계가 추가되거나 삭제된다. 예를 들어, 온라인 쇼핑이 소비자들에게 인기를 끌면서 전체 지출에서 차지하는 비중이 커짐에 따라 배송비와 현지 배달비가 2020년 바스켓에 추가되었다. 기초집계 수준에서 분류는 해당 분류의 모든 상품을 특성화하기 위해 선택된 품목(item) 표본을 포함한다. 대표상품은 널리

49) ①식료품(Food), ② 주거비(Shelter), ③ 가구 운영, 가구 및 장비(Household operations, furnishings and equipment), ④ 의류 및 신발(Clothing and footwear) ⑤ 운송(Transportation) ⑥ 보건 및 개인 서비스(Health and personal care) ⑦ 레크리에이션, 교육 및 독서(Recreation, education and reading) ⑧ 알코올, 담배 제품 및 대마초(Alcohol beverages, tobacco products and recreational cannabis)

구할 수 있고 소비자에게 가장 인기가 있다고 알려진 품목에 중점을 두고 선택되며, 선정된 품목이 실제로 구매하는 품목의 대표성이 되도록 한다. 기초집계에 할당되는 대표 상품(products) 수는 그 가중치와 더불어 해당 분류에 속하는 제품의 가격 변동성 및 이질성에 따라 달라질 수 있다. 예를 들어, 특정 건조 식료품 가격 측정시, 대표제품은 일반적으로 브랜드명과 매장 브랜드 품목을 모두 포함한다. 이와 대조적으로 바나나 집계 아래 가격이 매겨진 대표적인 제품은 단 한 가지뿐이다.

CPI 시계열 연속성은 바스켓에서 얻은 지수를 연쇄하고, 품목과 지역 각 집합 시리즈에 대해 별도로 수행된다. '20년 바스켓 가중치가 도입되면서 최근 소비 패턴변화를 반영해 새로운 제품 클래스가 추가됐다. 기본집계 '우편 및 기타 통신서비스'에는 두 개의 최하위 기초집계(배송료와 지역배달료)가 추가되었으며, 잡지 디지털 구독서비스가 기본집계 잡지와 정기 간행물에 추가되고, 비디오 게임기는 기본집계 비디오장비에 추가되어 2개의 기초집계가 추가되었다. 기초집계수준에서 12개 지수는 낮은 바스켓 비중 때문에 더 이상 간행되지 않는다(예: 발효 또는 절인 야채, 콘택트렌즈, 향수 등). 한 가지 눈에 띄는 내용은 '20년 바스켓에서 주류(2.94%)와 담배(1.27%)의 중요성이 높아졌고, 이는 SNA 데이터의 통합을 반영하며, SNA 자료는 자가 보고된 소비에 의존하지 않고 알코올, 담배 및 대마초에 대한 소비지출에 대한 보다 정확한 추정치를 제공한다는 것이다.

2. 미국 소비자물가지수

2.1 작성 방법 개요[50)]

두 인구 집단 각각에 대한 소비 패턴을 반영한다: 모든 도시 소비자와 도시 임금 근로자와 사무직 근로자이다. 모든 도시 소비자 집단은 미국 전체 인구의 약 93%를 차지하고, 임금 근로자와 사무직 근로자뿐만 아니라 전문직, 자영업자, 빈곤층, 실업자, 은퇴자를 포함한 도시나 대도시 지역

50) Monthly Releases for the Consumer Price Inde Technical note(BLS), Consumer Price Index Handbook of Manual(BLS, 11/24/2020)

의 거주자들 지출에 기초한다. 시골, 농가, 군대, 교도소, 정신병원과 같은 기관에 사는 사람들의 소비 패턴은 포함되지 않는다.

모든 도시 소비자의 인플레이션은 두 가지 지수, 즉 소비자 물가 지수(CPI-U)와 연쇄 소비자 물가 지수(C-CPI-U)로 측정한다. 도시 임금근로자 및 사무직근로자 소비자물가지수(CPI-W)는 다음 두 가지 요건을 충족하는 CPI-U 정의에 포함된 가구 지출에 기초한다. 가구소득 1/2이상이 사무직 또는 임금 직종이어야 하며, 가구소득자 중 적어도 한 명은 지난 12개월 동안 최소 37주 동안 고용되어 있어야 한다. CPI-W 인구는 미국 인구의 약 29%를 차지하며 CPI-U 인구의 부분 집합이다.

전국 75개 도시지역에서 매월 6,000여 주택과 약 2만 2,000여개의 소매업체에서 가격을 조사하고 있다. 품목구입 및 사용과 직접 관련된 세금이 지수에 포함된다. 연료와 기타 몇 가지 가격은 75개 모든 지역에서 매월 수집된다. 대부분의 다른 상품과 서비스 가격은 세 개의 가장 큰 지리적 영역((New York, Los Angeles, and Chicago)에서는 매월, 그리고 다른 지역에서는 격월로 수집된다. 대부분의 상품과 서비스 가격은 개별 방문이나 전화 통화를 통해 얻어진다. 지수 계산 시, 각 지역의 다양한 품목에 대한 가격 변동은 적절한 모집단 지출에서 그 중요성을 나타내는 가중치를 사용하여 집계된다. 그런 다음 지역 데이터를 결합하여 미국 도시평균을 구한다.

CPI-U와 CPI-W 경우, 도시규모, 지역별, 인구규모 및 23개 선택된 지역에 대해 지수를 생산한다. C-CPI-U는 전국지수만 생산된다. CPI-U와 CPI-W는 발표 시 최종이지만, C-CPI-U는 잠정 형태로 공표되며 이후 세 분기 개정이 따른다. 대부분의 CPI-U 및 CPI-W에서 기준값은 1982-84 = 100이다. C-CPI-U 기준치는 1999.12월 = 100이다.

표본오차 관련, CPI는 소매가격 표본에 기초하기 때문에 표본오차 영향을 받으며, BLS는 CPI-U의 매년 1개월, 2개월, 6개월 및 12개월 변동 표준 오류 추정치를 계산하여 발표한다. 이러한 표준 오차 추정치를

사용, 가설 검정을 위한 신뢰구간을 구성할 수 있다. 예를 들어, 1개월 변동 추정 표준 오차는 모든 품목 CPI의 0.03%이다. 모든 CPI-U 품목의 1개월 0.2% 변동의 경우, 95%는 모든 소매가격을 기준으로 한 실제 변동률이 0.14%~0.26% 사이가 될 것이라는 것이다.

지수 변동은 일반적으로 지수 포인트의 변화가 아니라 백분율 변화로 표현되는데, 이는 지수 포인트 변화는 기준 기간과 관련된 지수 수준의 영향을 받지만 백분율 변화는 그렇지 않기 때문이다.

2.2. 계절 조정[51)]

전국수준에서 계절변동의 중요한 패턴이 있는 CPI 구성요소에 대해 계절조정지수를 생산한다. 계절조정자료는 X-13 ARIMA-SEATS를 활용해 산출된 계절요인을 가지고, 계절조정자료는 매년 2월에 갱신되며, 새로운 factor들은 지난 5년간의 계절조정자료를 수정하는 데 사용 된다. 이 factor들은 홈페이지에서 확인할 수 있다. 경제 단기 물가 동향을 분석하기 위해 계절조정자료가 선호되는데, 이는 기후, 제품 생산 주기, 모델 변경, 휴일 및 세일로 인한 가격 변동과 같이 일반적으로 같은 시기에 거의 매년 동일한 규모로 발생하는 변화 영향을 제거하기 때문이다. 이를 통해 사용자(예: 연구자)는 연중 일반적인 변화가 아닌 변화에 집중할 수 있다.

원자료는 실제 소비자가 지불하는 가격에 대한 일차적 관심사이고, 에스컬레이션 목적으로 광범위하게 사용된다. 예를 들어, 단체 협상 계약과 연금 계획은 계절변동을 조정하기 전 물가 지수와 보상 변화를 연계한다. BLS는 계절조정 시리즈가 매년 개정되기 때문에 에스컬레이션 계약에서 계절조정 자료를 사용하지 않도록 권장한다.

일부 CPI 계열에 대해 개입분석을 사용한다. 극단적인 값(outlier)나 급격한 움직임(level shift)은 기조적인(underlying) 가격변동의 계절패턴을

51) BLS Consumer Price Index> Methods> Seasonal Adjustment

왜곡시킬 수 있다. 개입분석은 계절요인을 계산하기 전에 이러한 비정상적인 사건으로 인한 값을 추정하고 지수에서 제거(사전 조정)하는 것이다. 계절패턴을 더 정확하게 나타내기 위해 그 사전조정인자는 원계열에 적용된다. 예를 들어, 이 절차는 2008년 세계 경기침체 이후 2009년 가격 정상 복귀 효과를 상쇄하기 위해 모터연료에 사용되었다. 2022.1월 일부 식음료, 모터연료, 전기, 차량 등을 포함, 개입분석 계절조정을 이용해 72계열을 조정했다. 계절조정된 데이터는 최초 공표 후 최대 5년 동안 수정될 수 있다. 매년 CPI는 계절조정지수 산출을 위한 새로운 계절요인을 계산하여 최근 5년간의 데이터에 적용한다. 지난 5년을 초과한 계절조정지수는 최종 지수로 간주되며 수정의 대상이 아니다. 2022년 1월에는 2017년부터 2021년까지의 계절요인 및 계절조정지수를 계산하여 공표하였다. 매년 1월 모든 계열의 계절 상태는 특정 통계 기준에 따라 재평가되고, “계절조정되지 않음“에서 “계절조정“으로 또는 그 반대로 상태를 변경해야 하는지 여부를 결정한다. 미국 도시 평균 81개 구성요소 중 계절조정지수가 계절조정에서 계절 조정되지 않음으로 변경되는 경우, 계절 조정 않은 데이터는 지난 5년간 종속 계열의 집계에서 사용되지만, 그 기간 이전의 계절조정지수는 변경되지 않는다. CPI-U 품목지수 81개 중 22개는 2022년에 대해 계절적으로 조정된다.

2.3. 품질 조정[52)]

1) 가격조사 및 품질 조정

대부분의 가격조사는 조사자 방문을 통해 수행되지만, 다른 경우에는 웹사이트 방문이나 전화를 통해 수행된다. 샘플품목이 사용 가능한 경우 조사자는 가격을 기록하고 그 정보는 특정 재화나 서비스에 대한 자세한 지식을 가진 상품분석가가 검토한다. 이상한 가격 움직임을 검토하고 타당성을 점검한다. 물가지수 공식은 0(또는 무료)의 가격을 처리할 수 없기 때문에 0가격은 매우 작은 가격으로 조정된다. 샘플품목을 더 이상 이용할 수 없거나 최근 가격 수집 이후 상품 또는 서비스의 품질 또는

52) BLS Consumer Price Index> Methods> Quality Adjustment, Consumer Price Index, Handbook of Manual(BLS, 11/24/2020)

양(예를 들어 64온스 대신 59온스 오렌지 주스)이 변경된 경우 조사자는 이전 품목과 유사한 새 품목을 선택하는데, 이를 대체(substitution)라고 한다. 대체가 발생한다는 것은 상품분석가가 신 품목과 가격을 검토하는 것으로 그 새로운 가격은 지수 계산에 사용하기 위해 품질이 조정될 수 있다. 개념적으로, CPI는 동일한 품질 측정을 추구하지만, 품질 변화를 정확하게 수량화하는 것이 항상 가능한 것은 아니다.

2) 품목교체(item replacement) 및 품질조정

물가지수 산출시 가장 어려운 문제 중 하나는 제품사양과 소비패턴 변화에 따른 품질변화의 정확한 측정과 처리다. CPI는 변하지 않고 일정한 품질의 재화와 서비스를 구매하는 데 드는 비용을 측정해야 한다. 현실에서는 제품이 사라지고, 새로운 버전으로 대체되고, 신제품이 등장한다. 조사자는 CPI 샘플가격을 더 이상 얻을 수 없다는 것을 알게 되면(종종 아울렛이 영구적으로 판매중단) 신 품목을 찾기 위해 품목교체 절차를 사용한다. CPI 각 가격 품목계층은 하나 이상의 ELI(entry-level item)를 포함한다. CPI 분석가는 각 ELI의 추가 세분화를 정의하는 체크리스트를 개발 했고, 소매점에서 대체품을 구할 때, 조사자는 먼저 ELI 체크리스트를 이용하여 기존 가격이 측정된 품목에 가장 근접한 아울렛에서 판매하는 품목을 찾는다. 그런 다음 체크리스트에 교체품목을 설명하고 중요한 사양을 캡처한다. ELI 분석자는 품질변경 및 품목 사양 변경을 고려하기 위해 다음 세 가지 방법 중 하나를 선택한다.

다음 예에서 일반적 유형의 품질 조정을 설명한다. 조사자가 t기 품목 j의 가격을 수집하려고 하지만 해당 대상처가 더 이상 해당 품목을 팔지 않기에 조사를 할 수 없다고 가정하자(j의 가격은 t-1 기간에 수집). 조사자는 대체 품목을 찾아 그에 대한 가격을 수집한다. 이 교체 품목은 상품 j의 새 버전 v+1이 된다. 분석가는 대체품을 어떻게 취급할지 결정하는데, 상품 j의 두 가지 버전에 대한 설명을 가지고 있다. 이전버전 v의 t-1 가격 $P_{j,t-1}^{v}$, 대체버전 v+1의 t 가격 $P_{j,t}^{v+1}$이 있고, 다음 표는 분석가가 사용할 수 있는 정보이다.

버전	t-1 가격	t 가격
구버전 v	$P_{j,t-1}^{v}$	-
교체 버전 v+1	-	$P_{j,t}^{v+1}$

$P_{j,t-1}^{v+1}$, $P_{j,t}^{v}$ 둘 중의 가격이 없다면 품목 j에 대한 관찰은 기간 t에 대한 지수계산에서 제외되며, 이는 j품목 관찰이 무응답으로 처리된다. 분석가가 대체를 다루기 위해 선택할 수 있는 세 가지 방법은 다음과 같다.

① 직접 비교(direct comparison) : 품질차이가 없다고 가정

원래 품목과 대체품목이 본질적으로 동일할 경우 직접 비교할 수 있다고 간주하고 품목 간 가격 비교를 지수에 사용한다.

② 직접 품질조정(direct quality adjustment) : 품질차이 추정

품질 차이가 있는 대체품목을 다루는 가장 명확한 방법은 차이의 가치를 추정하는 것이다. 이 값의 추정치를 품질 조정량 $QA_{j,t-1}$이라고 한다. 이 경우, $P_{j,t-1}^{v+1} = P_{j,t-1}^{v} + QA_{j,t-1}$

$P_{j,t-1}^{v}$ 이전 버전의 t-1가격, $P_{j,t-1}^{v+1}$ 대체버전 v+1 기간 t-1가격

직접 품질조정은 크기나 무게, 제조업체 비용, 헤도닉 모델 등 관측 가능한 요소를 포함한다

③ 대체(imputation)방법

대체는 누락정보를 처리하기 위한 절차이고 응답자거부, 계절이 지났거나 다른 이유로 사용할 수 없는 품목, 품질변화 추정 불가능을 포함한 여러 사례에 대해 사용한다. 직접 비교하거나 품질이 조정될 수 없는 교체품목은 비교 불가라고 하며, 이 경우 품질 가격 변화 추정치가 만들어진다. 셀 상대 대체와 클래스 평균 대체 두 가지가 사용 될 수 있다.

· 셀 상대 대체(cell-relative imputation)

한 품목의 가격변동이 그 기본지수 안에서 다른 품목의 관측된 가격변동과 다르다는 이유가 없다면, 셀 상대법은 그 변동을 대체시키기 위해 사용된다. 이 방법은 결측값을 위해 사용되고, 원 품목과 비교불가능한 대체 품목 가격 변동은 동일한 지리적 영역에 대해 1개월 동안 유사 품목의 평균 가격 변동과 동일하다고 가정한다(즉, 해당 ELI 및 PSU의 기본 셀의 평균 가격 변동과 동일)

· 클래스 평균 대체(class-mean imputation)

일부 상품 및 서비스 품목 계층은 비교가 안 되는 교체품, 주로 차량, 기타 내구재 및 의류에 대하여 클래스 평균 대체를 사용한다. 이 논리는 가격 변화가 많은 품목에 대한 새로운 라인 또는 모델의 연간 또는 주기적인 도입과 밀접한 관련이 있다는 것이다. 예를 들어 새로운 연식 차량을 도입시 가격 상승이 종종 있는 반면, 연식 후반에는 가격 하락이 일반적이다. CPI는 제품 라인이 업데이트될 때 품목 교체를 처리하기 위해 가능한 한 자주 품질조정 방법을 사용한다. 클래스 평균 대체 방법은 나머지 교체상황에서 적용된다. 이 경우, CPI는 품목 교체를 거치며 품질을 직접조정 또는 직접 비교할 수 있다고 판단된 다른 관측치의 가격 변동으로부터 가격 변동을 추정한다. 클래스 평균 대체를 위해 구 버전 v의 현재 기간 t가격 추정치인 $P_{j,t}^{v}$를 추정하며, 이 추정된 현재 가격을 t기간의 상대 가격 계산에 사용한다. $P_{j,t}^{v} = P_{j,t-1} \cdot cR_{t,t-1}$ 추정 현재 가격은 구 버전 이전 기간 t-1 가격에 클래스 cR (특별히 구성된 상대 가격)을 곱한 것이다. 여기서 $cR_{t,t-1}$은 품목 j를 포함하는 ELI 관측치들의 부분집합에 대한 기하평균 또는 라스파이레스 공식으로 계산된다. 부분 집합은 관심 class로, 동일한 ELI 및 PSU(primary sampling unit)에서 모든 비교 가능하고 품질 조정된 교체 관측치들이다.

3) 품질조정 헤도닉

헤도닉 품질 조정은 이전 품목 가격에서 그 변동의 추정가치를 더하거나 빼서 품질의 변화로 인한 가격 차이를 제거한다.

① CPI : 품질변화를 위해 가격 조정하는 이유

CPI 측정 근본 문제는 새로운 버전 품목을 도입하고 이전 버전을 중단함에 따라 가격뿐만 아니라 그들 특성이 시간이 지남에 따라 변한다는 것이다. 새로운 버전은 추가 편익을 제공하거나 경우에 따라 편익을 줄일 수 있는데, 이 편익 변화가 품질 변화이다. 가격 변동을 정확하게 측정하기 위해서는 이 품질 변화로 인한 가격 변동분을 구별할 수 있어야 한다. 이 문제의 전통적인 해결책은 품질이 변경되었을 때 일시적으로 샘플에서 품목을 제거하는 것인데, 이는 때때로 허용 가능하지만, 새로운 버전의 가격 변동이 기존 상품의 가격 변동과 체계적으로 다를 경우 CPI를 편향시킬 수 있다.

② 헤도닉 품질조정

헤도닉 품질조정은 일부 CPI 샘플 내에서 품질변화를 설명하기 위해 사용하는 기법 중 하나이다. 제품 특성이 혁신이나 완전 신상품 도입으로 바뀔 때 가격을 조정하는 방식이다. 제품을 특성으로 분해하고, 각 특성에서 파생된 효용 추정치를 얻고, 상품 품질이 변경될 때 그 가치 추정치를 사용하고, 회귀분석을 통해 가격을 조정하는 데 사용되는 값 추정치를 얻는다. 헤도닉 회귀모델은 제품을 구성하는 각 특성에서 파생된 효용 값을 결정하기 위해 추정하는 것이다.

③ 헤도닉 : 품질변화 가치 추정 방법

샘플 각 상품에 대한 가격과 설명을 수집한다. 헤도닉을 사용하는 품목 범주에서 BLS는 회귀모델링을 사용하여 각 특성에 대한 값을 추정한다. 예를 들어 남성 셔츠 헤도닉 모델에서 소매 길이는 추정된 특성이고, 긴소매에 대한 추정치는 셔츠 가격 부분이다. 헤도닉은 시장에서 새로운

혁신을 포착하거나 기존 특성의 가치 추정에 변화를 반영하기 위해 약 2년마다 추정한다.

④ 헤도닉 회귀모형을 사용하여 품질조정 추정 및 적용하는 방법; 메커니즘을 설명하기 위해, 헤도닉 회귀 방정식의 일반화된 형태로 시작

$$\ln P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

종속변수 lnP가 가격 자연 로그이고, β는 독립변수(X_k) 계수추정치이며, ε는 오차항, 계수는 특성의 단위 변화와 관련된 가격의 비례적 변화의 측정치이다. 모델링 품목이 남성 셔츠인 경우, 독립변수는 소매 길이와 섬유 구성이고, 남성 셔츠에 대한 헤도닉의 단순화된 형태는 다음과 같을 수 있다.

$$\ln P = \beta_0 + \beta_1 (Longsleeve) + \beta_2 (100\% Cotton)$$

여기 모든 셔츠는 짧은 소매 또는 긴 소매와 면/폴리 또는 100% 면이고, 통계 처리를 수행한 후 β_1= 0.15 및 β_2 = 0.25로 추정할 수 있고, 긴 소매 셔츠가 짧은 소매 셔츠보다 15% 더 가치 있고 100% 면 셔츠가 면/폴리 합성 셔츠보다 25% 더 가치 있다는 것을 나타낸다. 샘플의 반팔 면/폴리 셔츠를 100%면 긴 소매 셔츠로 교체할 경우 신품목 특징에 따라 기존 품목의 가격을 조정해 약 49%($e^{0.15+0.25}$)의 가격조정으로 산출한다. 기존 셔츠 가격이 20달러이고 대체 셔츠 가격이 30달러였다면, 10달러 가격 인상을 사용하지 않고, 소매와 면 함량에 대한 기존 셔츠를 조정하여 \$29.84($20*e^{0.15+0.25}$)의 가격을 추정하게 된다. 셔츠 간의 가격 차이 대부분을 특성 변화로 기인하고 0.16달러 가격 상승으로 보는 것이다.

⑥ CPI에서 어떤 항목들이 헤도닉 방법으로 조정되었는가?

의류와 같이 계절 변화 또는 가전・전자제품과 같이 혁신적 개선과 기술 변화로 인해 높은 수준의 품질 변화를 경험하는 경향이 있는 품목 범주에서 사용한다. 해당 품목들 목록은 BLS 홈페이에서 확인가능하다.

Ⅴ 결론 및 시사점

1. 스캐너 데이터 활용한 CPI 작성

유럽 국가들과 그 외 캐나다, 호주, 뉴질랜드 등을 중심으로 빅데이터를 활용한 소비자물가지수 작성이 급속히 확산되고 있으며 관련 방법론에 대한 연구도 활발히 진행되고 있다. 스캐너 자료를 적용하는 국가들을 보면, 주로 슈퍼마켓 스캐너에 적용하고, 고정 가중치 접근법을 따르는 국가들과 변동가치 접근법을 수행하는 국가들이 있다. 캐나다, 네덜란드, 호주, 캐나다 등 16개국은 이미 스캐너 데이터를 활용해 CPI를 작성하고 있고, 네덜란드, 벨기에, 룩셈부르크, 노르웨이, 호주 등은 다변지수 산출 방식을 활용하고 있다.

우선, 통계청이 스캐너 자료를 이용하려면 안정적으로 자료를 제공받아야 한다. 통계법(예: 프랑스)을 적용하거나, 구매, 업무협약 등을 통해 이 데이터를 확보할 필요가 있다. 현재 한국의 경우, 민간 자료 수집에 대한 법적 근거 및 민간 부문의 자료 제공 유인 및 협조 부족으로 인하여 스캐너데이터 확보가 어려울 수 있다. 이에 법적근거 마련 및 MOU 체결 등 민간기업과의 협조체계 구축을 통해 안정적인 자료수집 방안을 지속적으로 모색할 필요가 있다.

스캐너 자료를 사용한 다른 국가들 경험은 소매업으로부터 직접 데이터 세트를 얻는 것이 더 바람직하다는 것을 시사한다. 단, 소매업체에서 직접 자료 세트를 확보할 수 없거나 양자 계약을 협상할 수 있는 자원이 부족한 경우 시장조사업체로부터 데이터를 얻는 것을 고려해 볼 수도 있을 것이다.

통계청이 스캐너 데이터를 확보한 경우, 지수를 작성하는 데 효과적이고 효율적으로 사용할 수 있는 정보로 변환해야 하고, 이를 달성하기 위해서는 IT 시스템을 개발하고, 데이터를 분류하고, 품질확보를 해야 한다. IT 시스템은 변화하는 방법 및 증가하는 데이터 공급자로부터 수신

되는 대량데이터에 맞춰 확장할 수 있도록 설계되어야 한다. 통계청은 이런 데이터를 본격적으로 사용하기 전에 이런 자료를 어느 정도 경험하는 것이 중요하다. 데이터 품질확보를 위해서는 글로벌 및 상세검사를 일상화해야 하고, 이러한 점검은 데이터 공급자와 접촉하고, 다른 가격 정보 소스(예: 전단지와 온라인 가격)와 비교할 수도 있다. 최종 집계 지수를 검토하고 신뢰성을 보장해야 한다.

분류 및 개별 제품(품목) 정의는 최종 결과에 상당한 영향을 미칠 수 있는 중요한 단계이다. 분류 관련해서, 분류체계의 신뢰성을 지속적으로 모니터링 할 필요가 있다. 이 단계에서 발생한 오류는 잘못 분류된 품목에 기초하여 하위 지수에 반영될 것이다. 이는 각 소매업체마다 제공하는 데이터가 다르고 상품코드 분류 역시 다를 수 있다. 초기 설정 단계에서, 어떤 상품코드가 사용되어야 하고 그 코드가 얼마나 안정적이어야 하는지를 상품 코드 사용시 고려해야 한다. 개별 제품(품목)은 시간, 대상처 및 제품 차원에 걸친 집계를 고려하여 신중하게 지정되어야 한다. 시간 범위는 가능한 한 많은 월 범위를 포함하는 기간에 걸쳐 종합하는 것이 좋다. 대상처 집계는 경험적인 문제이다. 개별제품은 상세한 대상처 수준에서 명시 되어야 하나, 데이터가 더 집계된 수준에서 제공되거나 가격 수준이 아울렛에서 유사한 경우(예: 동일한 유형과 체인) 아울렛 전체 집계가 고려될 수 있다. 재출시 및 제품 이탈은 나가는 품목 코드와 들어오는 품목 코드를 연결하여 처리될 수 있다. 재출시 및 제품 이탈에 대처하는 다른 전략은 품목코드 그룹핑을 만드는 것이다. 앞에서 의류 MARS 방법에서 제시한 바와 같이 균질한 제품을 만들 때 단위 값 편향과 시간 경과에 따른 매칭 모두 고려되어야 한다. 헤도닉 기법은 또한 제품 코드들 간의 전체 집계(동질제품)에 대한 대안이 될 수 있다.

산출방법은 한 가지만 있는 게 아니다. 대면 현장수집에 더 근접한 산출방법(고정 가중치 기반 직접 지수 변형)을 사용하는 국가들도 있고, 다른 나라들은 더 큰 표본과 적시 가중치를 갖는 접근법을 채택한다. 동적 바스켓 방식도 사용되고 있으며, 일부 국가들은 다변지수로 전환하기

시작하고 있다. 지수 대표성 편의를 줄이기 위해, 적절한 가중치를 사용하는 것이 가중치를 사용하지 않는 것보다 더 좋고, 시간 경과에 따른 제품 다양성을 포착하기 위해 샘플을 더 자주 업데이트하는 것이 더 유용하다고 볼 수 있다. 그런 점에서 다변지수 방법은 나름대로 기술적 어려움이 있지만 좋은 방안일 수 있다.

영국 통계청(ONS) CPI는 2023년부터 스캐너 데이터, 웹 스크래핑을 기존 데이터 수집 방법과 통합을 위해 연구 개발 중에 있다. 이 연구에서 스캐너 자료 처리 방법이 제품 분류, 시간 적용 범위, 고유 및 재출시 제품 식별・추적, 그리고 제품 크기 변화를 고려하여 가격 도출, 할인 처리 등을 포함하여 상점에서 대면으로 수집하는 데이터 처리와 어떻게 다른지를 검토하고 있고, 산출방법 품질프레임워크를 만들어서 스캐너와 웹스크래핑 데이터를 사용할 때 다른 국가 통계기관에서 사용 중인 다변지수를 포함하여 최저 수준의 집계에서 물가 지수를 산출하는 데 사용할 수 있는 여러 방법에 대한 테스트를 수행하였다. 결과는 최저 수준에서 웹 스크래핑 및 스캐너 데이터에 대해 다변지수 적용 CPI가 고정 또는 체인 양변지수 보다 더 포괄적이고 정확할 것이라는 것을 보여주었다. 다변지수를 사용하면 제품 적용 범위가 넓어지고 제품 수준에서 가중치를 매길 수 있으며 동적 데이터세트에서 수집된 제품 정보를 더 잘 활용할 수 있다. 또한 가중치와 가중치가 없는 방법 간의 차이가 더 컸다. 이는 웹 스크래핑 및 스캐너 데이터에 대해 제품 수준에서 판매 가치에 대한 정보 또는 근사치를 사용하는 것의 중요성과 이 연구가 가중 지수 방법과 확장 방법 그 자체의 선택보다 더 가치가 있다는 점을 보여주었다.

기존의 작은 샘플 고정 바스켓에서 다변지수와 같은 보다 발전된 기술에 이르기까지 방법의 복잡성이 증가하고 있다. 실무적으로 국가통계기관은 단계별 접근법(예 : 동적바스켓→ GEKS)을 적용할 수 있다. 시간이 지남에 따라, 전문성이 증가함에 따라, 더 나은 방법을 구현하는 것을 선택할 수 있다. 복잡성이 증가하면 사용자와의 커뮤니케이션에도 어려움이 발생한다. 소비자물가지수 측정의 기초가 되는 원칙과 방법은

투명하게 전달되어야 한다. Eurostat 및 UN 등 국제기구를 중심으로 스캐너 데이터를 활용하는 여러 방법을 논의하고, 제시하고 있다. 한국 통계청도 준비를 하고, 이러한 논의에 적극적으로 참여해야 할 것이다.

2. 웹스크래핑을 활용한 CPI 작성

웹 스크래핑은 웹에서 정보를 수집하고 적용하는 프로세스이다. 현장에서 수집되었던 가격을 공식 웹사이트에서 공개적으로 이용할 수 있는 데이터로 대체할 수 있는 여러 가지 이점이 있다. 이 접근방식은 기업의 응답 부담을 줄여 기업에 시간과 리소스를 절약하는 동시에 비용 효율적인 방식으로 고품질 데이터를 지속적으로 제공한다. 또한 대량의 정보를 더 빈번하게 입수하고 시기적절하고 정확한 통계를 작성할 수 있는 효율적인 수단이다.

현재 웹스크래핑을 적용하는 국가들은 의류 및 신발, 가전제품, 중고차, 열차, 항공료, 이동전화 등 특정제품 범주에 대해 적용하고 있으며, 가장 일반적으로 보고된 방식은 비가증 기하평균(제본스) 산식을 사용하지만 일부는 가중 기하평균도 사용하고 있으며 일부는 헤도닉 방법을 사용하고 있다. 현재 한국 통계청에서는 온라인가격 정보를 수집[53]하고 있지만, 물가지수 생산에는 직접 활용하지는 못하고 있다. 향후 웹스크래핑을 직접 활용하기 위한 Eurostat 사례에 비추어 기술적 측면의 고려사항을 살펴보고자 한다.

인터넷으로부터의 자료 수집과 통계 작성을 위해서는 전통적 데이터 수집 방식에서, 새로운 기술과 작업 프로세스 재편이 필요하다. 통계청 CPI 웹 스크래핑 프로젝트를 장기적으로 누가 책임지고, 어떻게 지속가능한 방식으로 정리할 것인가가 초기에 내려야 할 가장 중요한 결정일 것이다. 이는 대개 가용 인적과 재정 자원의 기능이기 때문에 더 많은 수의 IT 자격을 갖춘 직원이 있으면 프로세스가 더 쉬워질 것이

53) 온라인물가격정보는 웹페이지 URL을 구성하는 HTML(HyperText Markup Language)을 수집하는 기술인 '웹 스크랩핑(web scraping)' 기술을 이용하여 온라인 쇼핑 웹사이트에 제시된 모든 상품의 가격을 수집하여 작성되고 있으며, 기존의 소비자물가지수와는 포괄범위, 품질조정 등에서 차이가 있어 직접적인 비교는 할 수 없지만, 보조지표로 활용할 수 있다.(출처 : 통계청 행정편람 2020)

분명하고, 스크레이퍼를 내부적으로 코딩하면 더 나은 제어 옵션을 사용할 수 있다. 이는 훨씬 더 사용자 정의가 가능하며, 더 많은 해결 방법을 제공하고 제약도 줄일 수 있다. 그러나 웹스크래퍼 원 개발자들이 사업을 떠날 경우 등 생산을 위한 개발과 유지보수를 누가 담당하느냐이다 이상적인 상황은 CPI 직원과 긴밀히 협력 할지라도 스크래퍼 유지보수를 IT 부서가 맡는 것이다.

온라인 가격 수집에서는 웹 스크래핑을 목표가 아니라 최후의 선택으로 보는 것이 좋다. 통계청에서는 데이터 소유자에게 API에 대한 접근을 제공하도록 요청하는 것이 좋다. 이는 웹 사이트에 비해 더 안정적인 데이터베이스 사용을 수반하기 때문에 훨씬 더 확실한 기술 솔루션을 가능하게 한다. 잘 알려진 예는 수천 개의 항공료를 자주 수집할 수 있는 아마데우스 API이고, 이것의 활용은 좋은 출발점이 될 수 있다.

또한 웹사이트 운영자와 사전에 연락을 구축하는 것이 좋다고 이미 언급한 바 있다. 소매업체가 통계청에게 실제 판매량과 수량에 대한 정보를 제공하는 데 더 적합한 스캐너 데이터를 제공하는 데 동의할 수 있으므로 보다 효율적인 해결책으로 이어질 수 있다. 그들은 데이터를 가장 잘 알고 있으며 시스템에 대한 액세스 권한, 웹 사이트 정책 관련 정보, 웹사이트 변경시 사전 정보 및 가격 책정 알고리즘에 대한 통찰력까지 제공할 수 있다. 또한 차단을 피하기 위해 많은 협상이 필요하다면 스캐너 데이터를 얻기 위해 직접 노력을 투자하는 것도 고려해볼 만하다.

내부적으로 코딩하여 스크래핑을 개발하고 유지하는 것이 가장 좋은 옵션이다. 이는 웹페이지 열기, 클릭, 스크롤과 같은 모든 웹브라우저 상호 작용을 코딩(또는 스크립팅)함으로써 자동화할 수 있으며 정확한 쿼리를 실행하여 관심 요소를 추출할 수 있다. 일반적으로는 코딩하면 필요에 맞게 데이터를 더 잘 처리할 수 있고, 장기적으로 더 유연하고 저렴하다. 그러나 웹사이트 변경은 많은 오작동을 일으킬 수 있으므로 스크래퍼를 개발하고 사용하는 통계청에서 사용자 친화적인 프로그램을 개발하지 않는 한 프로그래밍 기술을 갖춘 인력과 이를 실행할 서버가

필요하다. 대부분의 사이트들이 맞춤형 스크래퍼를 필요로 하며, 웹사이트의 작은 변경에 대해 유지되어야 하기 때문에 각 사이트마다 전용 스크래퍼를 설치하는 것은 비용이 많이 들 수 있다. 그렇기에 점진적으로 시행하고, 보다 안정적이고 통제가 복잡하지 않을 것으로 예상되는 웹사이트를 선정하고, 다루기 쉬운 품목(item)부터 시작하는 것이 좋다.

웹스크래핑 활용을 결정할 때는 컴파일과 지수계산 방법뿐 아니라 직원들에게 가장 적합한 웹스크래핑 실행이 무엇인지도 결정해야 한다. 이른바 표적(target) 스크래핑 수행시 컴파일, 지수 계산 방법을 그대로 수행하고, 실제 가격 수집만 자동화된다. 따라서 웹 스크래핑 프로그램은 인간 가격 수집가의 행동을 모방한다. 이 방법이 웹 스크래핑 사용을 시작할 때 권장된다. 또한, 스크래퍼 사용자는 코드 구문을 이해하고 웹 사이트 변경 시 필요한 스크래퍼 파라미터를 업데이트할 수 있을 정도로 교육을 받는 것이 중요하다. 예를 들어 구조노드(structure nodes) 차이는 일반적으로 더 간단한 방식으로 조정할 수 있지만, URL 변경이나 텍스트에서 그림으로 데이터 표시변경은 스크래퍼를 조정하는데 더 큰 문제를 일으킬 수 있다.

3. 캐나다 및 미국 소비자물가지수 작성 방법

현재 한국 통계청은 계절조정자료를 생산하고 있지 않다. 그러나 캐나다 및 미국은 계절조정지수를 생산하고 있으므로, 한국도 적용여부를 검토해 볼 필요가 있다. 경제의 단기 물가 동향을 분석하기 위해 계절적으로 조정된 자료가 선호되는데, 이는 기후, 제품 생산주기, 모델변경, 휴일 및 판매로 인한 가격 변동과 같이 일반적으로 같은 시기에 거의 매년 동일한 규모로 발생하는 변화의 영향을 제거하기 때문이다. 이를 통해 데이터 사용자는 연중 일반적인 변화가 아닌 변화에 집중할 수 있다.

원자료는 실제 지불가격에 대한 이용자의 일차적인 관심사이고, 에스컬레이션 목적으로 광범위하게 사용된다. 예를 들어, 많은 단체협상 계약과 연금 계획은 계절적 변동을 조정하기 전에 소비자물가지수와 보상변경을 연계한다. 캐나다 및 미국은 계절조정치는 매년 개정되기 때문에

에스컬레이션 계약에서 계절조정 데이터를 사용하지 않도록 권장한다.

캐나다 통계청은 전국 및 지역 수준에서 물가지수(원계열)를 생산하고 계절조정지수는 전국수준에서 총 지수, 8개의 주요 분류 및 4개의 특수분류 전체 13개 계열을 산출하고 있다. 계절조정계열은 X-12-ARIMA를 사용하여 산출된다. 미국 BLS와 달리 계절조정 프로세스는 각 계열이 직접 조정되고 하위 구성요소를 집계한 결과가 아니다. 계절조정지수는 수정 가능한 유일한 CPI 계열이고, 매월 전월 계절조정지수가 수정 대상이다. 지난 3년간 계절적으로 조정된 값은 매년 1월분 공표시 수정된다.

미국 BLS는 전국 및 지역 수준에서 물가지수(원계열)를 생산하고 계절조정지수를 전국 수준에서 선택된 그룹과 하위 그룹에 대해 생산한다. CPI-U 총지수와 그 하위 집계의 계절변동은 계절조정된 구성요소 지수를 집계하여 도출된다. 매년 1월 모든 지수 계열의 계절조정 상태는 특정 통계 기준에 기초하여 재평가된다. 지수는 계절조정 상태를 계절조정에서 계절조정되지 않음으로 변경하거나 그 반대로 변경할 수 있다. 매년 2월에 1월 데이터가 공개되면 보정된 계절조정지수가 발표되며, 계절조정인자는 사용자가 요청할 때 이용할 수 있다. 계절조정은 X-13ARIMA-SEATS를 사용한다. 개입분석 및 데이터 확장을 위해 regression-ARIMA과 X-11을 함께 사용한다. 일부 CPI 계열에 대해 개입분석을 사용하는데, 이는 극단적인 값(outlier)이나 급격한 움직임(level shift)이 근본적인 계절적 가격 변동 패턴을 왜곡시킬 수 있기 때문이다. 미국 도시 총지수를 포함하여 계절조정지수는 최초 공표 후 최대 5년 동안 수정될 수 있다. 매년 BLS는 새로운 계절 요인을 계산하여 최근 5년간의 데이터에 적용한다. 지난 5년간을 초과, 그 이전 계절조정지수는 최종 지수로 간주되어 수정 대상이 아니다.

캐나다 통계청 가증치 출처는 기본등급 가중치는 SNA 가계최종소비지출에 기초하고 하위수준(기본등급 아래) 가중치는 다른 출처(스캐너 데이터 등)에서 얻을 수 있다. 현재 가중치 기간은 2021년 연간 지출 기준이며, 기본등급 범주이하 가중치는 언제든지 갱신 가능하다. 가중치 업데이트 빈도는 2년 주기이며 연간 갱신 가능성을 연구 중이다. 최근 Covid로

인해 2021년 1월에 계획되었던 개편은 당해년 7월로 연기되어 수행되었고, 2022년 6월에 2021년을 기준으로 한 가중치 갱신이 이루어졌다.

캐나다 통계청 및 미국 BLS 모두 CPI에 연동되는 지급 및 계약의 변경을 방지하기 위해 CPI를 개정하지 않고, 비개정 관행은 대부분 국가통계기관의 일반적 관행과 일치하며 ILO가 채택한 CPI 결의안에서도 확인되었다. 이를 위해 CPI 기본지수를 상위레벨로 집계하는 데 고정가중치(바스켓)를 적용하는 Lowe 산출 방법을 적용하고 있다. 2년 고정 바스켓[54] 간 연쇄는 바스켓 갱신 월 시점(월 중첩)에서 이루어진다. 이를 위해 바스켓의 혼합지출 가중치를 공통기간가격(연결월)으로 표현한다. 연결 월 가격으로 표현되는 혼합지출을 얻기 위해 원래 지출 가중치를 가격 갱신하여 구한다. 현행 한국 통계청의 바스켓 간 공통기간가격이 연간이고, 가계동향조사 소비지출이 시차가 있으나 가격 갱신하지 않는다. 이런 차이로 수정문제가 발생한다.

우리나라 통계청은 2014년 국가통계위원회에 보고된 "2015년 기준 CPI 개편계획"을 통해 현행 중간년도 CPI 가중치 개편시 사용되고 있는 산출방식의 문제를 인식하고 과거 시계열이 수정되는 문제를 해결하고자 하는 방안을 제시하였는데, 그 방안은 가중치 상위단계 계산시 가격이 보정된 가중치와 연결(중첩)방법을 연간에서 월로 지수접속법을 적용하는 Lowe방식으로 C소비자물가지수를 작성하는 것이다. 이 방식 적용은 과거 시계열 수정 없이 지수를 작성할 수 있다. 이에 대한 문제점을 진단하고 개선해 나가야 할 필요성이 있다.

54) 최근 미국 BLS는 2023년 1월 자료를 시작으로 단일 연도를 기준으로 소비자물가지수에 대한 가중치를 매년 업데이트할 계획이다. 이는 2년간의 지출 데이터를 사용하여 2년마다 가중치를 업데이트하던 이전의 관행에서 변화를 반영한다. 내년에는 2021년 소비지출 자료를 사용할 것이다.

[참 고 문 헌]

1. 국내문헌

통계청(2017). 소비자 물가지수 생산을 위한 스캐너 데이터 활용방안 연구

통계청(2014). 소비자물가지수 개편계획(안), 경제통계2분과위원회 회의자료

2. 외국문헌

International Labour Office. (2020). “Consumer Price Index Manual: Theory and Practice.” Geneva.

영국 ONS, Research into the use of scanner data for constructing UK consumer price statistics April 2021

영국 ONS, New index number methods in consumer price statistics 1 September 2020

영국 ONS, Research and developments in the transformation of UK consumer price statistics: September 2020, 1 September 2020

Index Compilation Techniques for Scanner Data, Claude Lamboray, UNECE ,Online meeting, June 2021

Practical guidelines on web scraping for the HICP, Eurostat, November 2020

Guide on the use of multilateral methods in the HICP Draft version (October 2021)

Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data.

Chessa, A. G. (2018). MARS: A method for defining products and linking barcodes of item relaunches, NTTS 2019 conference, Brussels, Belgium.

UN Task Team on Scanner Data, Methods and Applications(2021.11.9.)

Scanner and Web Scraping Eurostat Workshop, Eurostat, 2021.10.

- From contact to data provider to reception of data, UN Task Team Scanner Data, Kristiina Nieminem, Federico Polidoro)
- Recent developments in Luxembourgish CPI: from dynamic basket to multilateral methods, STATEC, Luxembourg – Botir Radjabov

The Canadian Consumer Price Index Reference Paper(Statistics Canada(2019.2)

SDDS – DQAF View _ Canada – Price index_ Consumer prices, IMF

Prices Analytical Series : Enhancements and Developments in the Consumer Price Index Program(Statistics Canada, February 17, 2021)

Enhancements to the Air Transportation Index in the Consumer Price Index (Statistics Canada, January 22, 2020)

Statistics Canada Consumer Price Index: Frequently asked questions

The Integration of Web-Scraped Data into the Clothing and Footwear Component of the Consumer Price Index(Statistics Canada, February 19, 2020)

Consumer Price Index, Handbook of Manual(BLS, 11/24/2020)

Quality adjustment in the BLS programs, Data Users' Conference(2021.4)

Monthly Releases for the Consumer Price Index(BLS)

Timeline of Seasonal Adjustment Methodological Changes(BLS, 02/08/2021)