The Journal of Korea Research Society for Customs
Volume 26, Number 3, August 2026 : pp. ?~??.
관세학회지 제26권 3호 (2026년 8월 28일)

논문투고일자 : 2026. 5. 12.
심사(수정)일 : 2026. ??. ??.
게재확정일자 : 2026. ??. ??.

# 합성품목번호를 사용한 대한민국 HSK의 시계열 연계

## Time-Series Linkage of Korea's HS codes Using Synthetic ID

정 덕 재**

< 초 록 >

무역 통계의 분석은 HS 코드의 잦은 분화·수렴으로 인하여 시계열적인 단절 문제를 안고 있다. 이러한 한계를 극복하기 위해 획기적인 대안으로 제안된 것이 Pierce and Schott (2009)의 합성품목번호 기법이다. 그러나 저자가 조사한 바에 따르면, 국내외 논문 및 보고서에서 한국의 HSK를 대상으로 합성품목번호 기법을 적용한 사례는 존재하지 않았다. 본 논문은 이러한 공백을 메우기 위해 대한민국의 2010년부터 2026년까지 17년간의 기간 동안의 HSK 10단위의 전 품목을 대상으로 Pierce and Schott (2009)의 합성품목번호 기법을 적용하여 일관된 시계열 연계표를 구축하였다. 구축하는 코드는 Python 버전 및 STATA 버전으로 Github에 공개한다. 또한 시계열 연계 결과를 누구나 시각적으로 조회할 수 있도록 실행 파일(exe) 형태로 가공하여 동일 GitHub에 공개함으로써, 후속 연구자들이 한국 무역 데이터의 미시 시계열 분석에 참고할 수 있도록 하였다.

JEL 코드: C81, F14, F13, C55, F10
주제어 : 합성품목번호, 관세품목분류표(HSK), 시계열 연계표, Pierce-Schott 알고리즘

[목 차]



* 한국관세무역개발원 연구위원, 경제학 박사, 이메일: ubuzuz@gmail.com, 홈페이지: www.jayjeo.com

# Ⅰ. 서 론

국제 무역 데이터의 실증적 시계열 분석에 있어 가장 기초적이면서도 치명적인 오류를 유발할 수 있는 요인은 바로 품목분류코드(Harmonized System, HS)의 시간적 불연속성이다. 세계관세기구(WCO)가 관장하는 HS 코드는 전 세계 교역품을 통일된 기준으로 분류하기 위해 도입되었으나, 국제 무역 환경의 변화, 기술 진보에 따른 신제품의 등장, 그리고 교역량이 미미해진 기존 품목의 퇴출 등을 반영하기 위해 통상 5년 주기로 대대적인 개정을 겪는다. 나아가 각국 통계 및 관세 당국은 자국의 세부적인 무역 수지 파악과 관세율 적용을 위해 국제 표준인 HS 6단위를 세분화하여, 미국은 10단위의 Schedule B(수출) 및 HTS(수입) 코드를, 대한민국은 10단위의 HSK(Harmonized System of Korea)를 운용하며 이를 매년 혹은 수년에 한번씩 개정하고 있다.

이러한 품목 분류의 지속적인 갱신은 당대의 경제 현실을 정확히 반영한다는 장점이 있으나, 장기 시계열 데이터를 구축하여 산업의 진화나 무역 정책의 효과를 추정하려는 연구자들에게는 심각한 통계적 착시를 유발한다. 가장 빈번하게 발생하는 문제는 기존 품목 코드의 '수렴(Convergence)'과 '분화(Divergence)' 현상이다. 간단한 예시를 통해 이 문제의 본질을 파악할 수 있다. 2020년 기준으로 사과의 HSK 코드를 A, 오렌지를 B, 복숭아를 C라고 가정할 때, 2021년에 관세청이 이 세 품목을 모두 A라는 단일 코드로 통폐합(수렴)하고 B와 C를 폐지했다고 가정해 보자. 이 경우, 실질적인 사과의 수입량에 전혀 변화가 없었음에도 HSK A 코드의 2021년 수입 금액은 2020년 대비 급격히 상승한 것으로 집계된다. 오렌지와 복숭아의 교역액이 A 코드에 합산되었기 때문이다. 반대로 2019년에 A 코드 하나에 사과, 오렌지, 복숭아가 모두 포함되어 있다가 2020년에 각각 A, B, C 코드로 분리(분화)되었다면, 2020년 A 코드의 교역액은 전년 대비 폭락한 것으로 나타난다. 오렌지와 복숭아의 물량이 B와 C로 빠져나갔기 때문이다.

이처럼 단순한 연계 결과만을 바탕으로 특정 HS 코드의 수출입 금액, 중량, 단가를 시계열로 비교하는 것은 위험한 접근이다. 통계적으로 나타나는 교역액의 급증이나 급감이 실제 시장 수요의 변화나 무역 협정의 효과인지, 아니면 단순히 세관의 코드 분류 체계 개편에 따른 행정적 부산물인지 분간할 수 없게 되기 때문이다. 미국의 주요 국책 연구기관과 학계는 이러한 시계열 단절이 제품의 진입과 퇴출에 따른 외연적 한계(Extensive margin) 변화를 심각하게 과대 추정하게 만든다는 사실을 발견하였다. Pierce and Schott (2009) (이하 P&S라고 칭한다)이 자신들의 합성품목번호 연계표를 적용하여 산출한 통계에 따르면, 1989년부터 2004년까지 16년 동안 2004년 기준으로 미국 수입 품목의 49%, 수출 품목의 40%가 HS 코드의 변경을 한 차례 이상 겪었으며, 이렇게 코드 변경 이력이 있는 품목들이 같은 해 미국 수입액의 62%, 수출액의 48%를 차지한 것으로 나타났다. 이는 분류 체계의 시계열 단절을 통제하지 않을 경우 미국 교역액의 절반 가까이가 통계적 착시에 노출될 수 있음을 의미한다. 나아가 Bernard, Jensen, Redding and Schott (2009)은 이러한 P&S 연계표를 적용하여 1993~2003년 미국의 외연적 한계를 새로

측정한 결과, 미국 경기침체기(2000~2002년)를 제외한 단기 1년 변화 구간에서는 평균적으로 76%의 수출액 변화가 HS 코드의 시계열 단절에서, 25%가 품목·교역 상대국의 추가·중단에서 비롯되며, 기업의 진입·퇴출 자체가 차지하는 순 기여도는 미미한 수준임을 보였다. 즉 분류 체계 변동을 통제하지 않으면 미시적 외연적 한계가 과대 추정되지만, 통제 후에는 그 비중이 학계가 직관적으로 가정해 온 수준보다 훨씬 작다는 사실이 실증적으로 확인된 셈이다.

이러한 문제를 근본적으로 해결하기 위해 학계에서 현재 가장 대표적으로 통용되는 방법론이 바로 P&S가 고안한 '합성품목번호(Synthetic ID)' 기법이다. 이 기법은 분석 대상 기간 동안 수렴 및 분화의 과정을 거친 모든 HS 코드를 추적하여 하나의 거대한 '가족 트리(Family tree)'로 묶고, 여기에 고유한 합성품목번호[2]를 부여하는 방식이다. 앞선 과일의 예시에 이 기법을 적용하면, 사과, 오렌지, 복숭아가 수렴되었든 분화되었든 상관없이 이들이 교차된 이력이 존재한다면 2019년부터 2021년까지의 모든 기간에 걸쳐 이 세 품목을 하나의 합성 그룹으로 병합 처리한다. 그 결과 특정 연도에 품목이 합쳐지거나 쪼개지면서 발생하는 수입 금액의 비정상적인 급격한 증가나 감소 현상을 효과적으로 상쇄할 수 있다.

본 논문은 이러한 이론적 배경을 바탕으로, 대한민국의 2010년부터 2026년까지의 HSK 10단위 연계표를 P&S의 방법론을 활용하여 구축하였다. 이를 위해 국내 학술지 및 국책연구기관 보고서에서의 적용 사례를 조사하여 선행연구의 공백을 확인하고, 해외 학계에서 논의되어 온 HS 연계 기법의 발전사를 서술하며, 궁극적으로 P&S 기법이 지니고 있는 태생적 한계점과 이를 극복하기 위한 최신 연구 동향을 심층적으로 분석하여 향후 대한민국 HSK 분석 논문 작성을 위한 견고한 방법론적 토대를 제공하고자 한다.

# II. 선행연구

## 1. 대한민국 논문에서의 Pierce and Schott (2009) 적용 사례

우선 국내 KCI 등재지 및 국제 SSCI 저널에 게재된 한국 학자들의 논문에서 대한민국의 HSK를 대상으로 P&S 기법을 적용한 선행연구가 존재하는지 조사하였다. 조사 결과, 공개적으로 확인 가능한 범위 내에서 P&S의 합성품목번호 알고리즘을 대한민국의 고유한 통관 분류 체계인 HSK 10단위에 직접 프로그래밍하여 시계열 데이터베이스 자체를 구축하고 그 연계 결과를 주된 방법론적 성과로 보고한 학술지 논문은 확인되지 않는다.

국내 KCI 등재 학술지에서 HSK를 본격적으로 다루는 연구들은 존재하나, 그 초점이 시계열 연계 자체에 있지 않다. 예컨대 박윤수·전태완·신선경·엄남일 (2022)은 수출입 폐기물의 관리 개선 방안을 다루며 HSK 분류 체계 내에 폐기물 품목이 어떻게 위치하는지를 검토하고 신설 품

2) P&S에서는 이것을 setyear로 칭한다.

목 119개를 제안하였다. 김진규 (2022)는 제7차 HS 협약 개정에 따라 HS 제3907호의 PETG 품목분류 신설 문제를 사례 연구로 수행하였으며, 김중선·심상렬 (2022)은 한국 반도체 산업의 표준품목분류 체계를 마련하기 위해 산업통상자원부의 MTI 코드와 HS 코드를 비교 분석하였다. 그러나 이들 문헌은 모두 HSK 시계열 코드의 변동을 가족 트리(Family tree) 방식으로 추적하여 합성품목번호(Synthetic ID)를 부여하는 P&S식 문제 설정과는 본질적으로 다른 주제를 다루고 있다.

국내 학술지에서 HSK 10단위 코드의 시계열 연계 자체를 알고리즘적으로 구축한 사례가 부재한 이유는, 상당수의 실증 연구가 미국 통계국(US Census Bureau)이나 전미경제연구소(NBER) 등에서 이미 가공하여 제공하는 미국 중심의 SIC/NAICS-HS 연계 데이터베이스를 그대로 가져다 쓰거나, 광범위한 산업 수준(예: KSIC 3단위나 4단위)으로 무역 데이터를 대폭 집계(aggregate)함으로써 HSK 10단위 코드 수준에서 발생하는 미시적인 분화 및 수렴의 문제를 우회하는 방식을 택해 왔기 때문이다.

즉, 매년 한국 관세청이 고시하는 관세율표와 통계부호표를 기반으로, 특정 연도의 HSK 10단위 코드가 다음 연도의 어떤 코드로 소멸되고 신설되었는지 그 신구 대조표를 전수 수집하여 P&S의 수학적 가족 트리 알고리즘을 한국의 실정에 맞게 코딩하고, 이를 통해 산출된 17년간(2010~2026년)의 고유한 HSK 합성품목번호 연계표의 생성 과정 및 그에 따른 교역량 보정 효과를 논증한 학술지 논문은 공개적으로 확인 가능한 범위에서 확인되지 않는다. 따라서 본 논문은 한국 무역 데이터를 활용하고자 하는 수많은 후속 연구자들에게 유용한 연계표를 제공한다는 점에서 의의가 있다.

## 2. 대한민국 연구기관에서의 Pierce and Schott (2009) 적용사례

학계의 학술지 논문에서는 HSK 연계표 자체를 알고리즘적으로 구축하는 시도가 확인되지 않는 반면, 장기적인 시계열 무역 데이터를 바탕으로 특정 무역 협정의 경제적 효과를 엄밀히 평가해야 하는 대한민국의 국책연구기관 보고서에서는 P&S의 방법론을 도입하여 데이터의 불연속성을 통제한 구체적인 적용 사례가 확인된다. 가장 대표적인 기관은 대외경제정책연구원(KIEP)으로, 이들은 미국의 교역 데이터를 장기간 다루는 과정에서 해당 기법의 필요성을 절감하고 이를 실증 분석에 직접 반영하였다.

KIEP의 정재욱·김예진 (2018)이 발간한 연구보고서 「미국 아프리카성장기회법(AGOA)의 교역 효과와 정책적 시사점」은 P&S 기법이 국책 연구에 어떻게 실질적으로 응용될 수 있는지를 보여주는 모범적인 사례이다. 이 보고서는 미국이 사하라 이남 아프리카 국가들의 경제 성장을 지원하기 위해 2000년에 제정한 아프리카성장기회법(AGOA, African Growth and Opportunity Act)이 미국과 아프리카 국가 간의 실제 교역 증감에 어떠한 영향을 미쳤는지를 계량적으로 분

석하는 것을 주된 목표로 삼았다.

연구진은 분석을 수행함에 있어 치명적인 통계적 난관에 봉착했다. AGOA의 효과를 검증하기 위해서는 제정 초기인 2000년대 초반부터 최신 시점인 2017년까지 약 16년 이상의 양자 간 교역 시계열 데이터를 분석해야 했다. 그러나 이 기간 동안 세계관세기구(WCO)는 5년 주기로 상위 6단위 국제 공통 코드를 지속적으로 개정하였으며, 미국 내 통관 당국 역시 자국의 수출입 규정 변화 및 관세 부과 기준의 세분화 요구에 맞추어 HS 품목 코드 10단위(HTS)에 대한 조정을 연중 수시로 단행해 왔다. 이로 인해 교역 자료 내 품목 코드의 극심한 불연속성이 발생하였고, 이를 그대로 분석에 투입할 경우 단순한 코드 체계 개편으로 인한 특정 품목의 일시적 무역액 소멸이나 급증을 AGOA의 정책적 파급 효과로 오인할 위험이 매우 높았다.

이러한 방법론적 한계를 교정하기 위해 KIEP 연구진은 Frazer and Van Biesebroeck (2010)의 분석 모형을 차용함과 동시에, 시계열 데이터의 무결성을 확보하기 위한 핵심 장치로 P&S의 알고리즘을 도입하였다.[3] 연구진은 P&S가 제공한 부속 코드 변환표(2009년까지 유효)를 미국 ITC의 연간 코드 변화 내역을 참고하여 최근까지 연장한 뒤 미국 HS 6단위 코드에 적용하였다. 그 결과 1997년부터 2009년까지 HS 6단위 기준으로 코드 변화가 없는 174개 품목과 변동이 있었던 2,970개 상품군, 총 3,144개 상품군으로 재분류하는 데 성공하였다. 즉, 특정 연도에 품목이 세분화되거나 폐지되어 다른 코드로 흡수된 이력이 있다면, 이들을 모두 엮어 시계열 내내 변하지 않는 하나의 일관된 상품군 기준으로 통일한 것이다.

이러한 연계표 구축의 결과는 성공적이었다. 연구진은 통계적 왜곡이 제거된 정제된 데이터를 바탕으로, 외부적 경제 충격이나 아프리카 국가들의 거시적 교역 조건 변화를 적절히 통제한 상태에서 AGOA의 순수한 교역 증가 효과를 산업별, 시기별로 뚜렷하게 분리하여 확인해 낼 수 있었다. 특히 AGOA에서 핵심적으로 관세 특혜를 허용하는 아프리카산 의류 제품의 경우, 일반 수출 상품에 비해 미국으로의 교역 증가 효과가 압도적으로 크다는 점을 실증적으로 증명하는 성과를 거두었다. 만약 P&S 기법을 통한 연계 과정을 생략했다면, 의류 품목 내에서 빈번하게 일어나는 섬유 및 디자인에 따른 HS 코드의 잦은 분화 현상으로 인해 AGOA의 진정한 혜택 규모를 과소 추정하거나 통계적 유의성을 확보하지 못했을 가능성이 크다.

이 사례는 장기적인 무역 시계열 분석을 수행하는 국가 단위의 정책 보고서에서 품목 코드의 연동 문제를 해결하기 위해 P&S 기법이 얼마나 중요한 방법론적 무기로 활용되는지를 보여준다. 그러나 유의미한 점은, KIEP의 선행연구도 미국과 아프리카 간의 통상 관계를 규명하기 위해 미국의 ‘HS 6단위 국제 공통 코드’에 집중하여 동 기법을 적용했다는 사실이다. 즉, 대한민국

---

3) 다만 KIEP 보고서가 인용한 P&S의 표기는 동 저자들의 「Journal of Official Statistics」 게재본인 2012년판이나, 진짜 P&S 알고리즘의 원형은 본 연구가 기준으로 삼는 P&S인 “Pierce and Schott (2009), NBER Working Paper, WP No. 14837” 이다. P&S (2009)는 시간에 따라 분화·수렴되는 HS 코드를 가족 트리(family tree)로 추적하여 합성품목번호를 부여하는 HS-HS 시계열 연계 알고리즘을 다룬다. 반면, 동일 저자들의 P&S (2012)는 미국 HS 10단위 코드와 산업분류(SIC/NAICS) 사이의 매핑을 다루는 HS-산업분류 간 연계 논문이다. 두 논문은 제목에 모두 ‘Concordance’를 포함하고 동일 저자 조합을 공유하나, 전자는 시간 축에서의 코드 변동을, 후자는 분류 체계 간의 코드 변환을 다루는 근본적으로 다른 주제를 취급한다.

관세청이 관리하는 'HSK 10단위' 체계의 고유한 개정 연혁을 전수 조사하여 한국 경제 전용의 합성품목코드 연계표를 구축한 것은 아니었다.

## 3. 해외 선행연구 및 Pierce and Schott (2009)의 제안

미국이나 유럽의 학계 및 통계 발행 기관들은 HS코드의 수렴이나 분화로 인해 발생하는 시계열의 불연속성이 제품의 진입과 퇴출, 혹은 특정 산업의 무역 성장률을 심각하게 오판하게 만든다는 문제점을 오래전부터 인지하고 있었으며, 이를 해결하기 위해 지속적으로 연계 기법을 고도화해 왔다.

가장 고전적인 방식의 연계는 정태적인 일대일(1:1) 또는 다대일(N:1) 매핑 테이블을 구성하여 과거의 데이터를 강제로 새로운 체계에 이식하는 방식이었다. 대표적으로 Feenstra (1996)는 1972년부터 1988년까지 사용되던 미국의 구형 관세 분류 체계인 TSUSA(Tariff Schedule of the United States Annotated)를 새로운 HS 체계로 전환하기 위해 기초적인 연계표를 구축하여 학계에 제공하였다. 이와 유사하게 UN Comtrade는 서로 다른 연도 버전의 HS 코드를 연계하기 위해 후방 변환 기법을 사용해 왔다. UN Statistics Division이 자인하듯, 이 방식은 신(新) HS 분류의 단일 세번(subheading)에 구(舊) 분류의 단일 세번을 강제로 할당하는 방식으로, 분류 개정의 본질이 일대일 대응을 허용하지 않음에도 불구하고 임의로 1:1 관계로 처리한다는 본질적 한계를 갖는다. 그 결과 시계열 분석 시 전체 품목의 수가 인위적으로 급감하거나 특정 품목의 교역액이 극단적으로 쏠리는 왜곡이 지속적으로 발생하였다. Cebeci (2012)는 2002~2010년 기간에 59개국 HS 6단위 수출 자료를 사용하여, 비개정 연도에는 약 8~9% 수준이던 제품 퇴출률이 2007년 HS 개정 시점에 자가신고 데이터에서 약 15%, Comtrade의 단순 변환 데이터에서도 약 12%까지 일시 급등하는 현상을 실증적으로 보였다. 즉 단순 후방 변환은 개정 연도마다 통계적 단절(spike)을 그대로 노출시킨다.

이러한 정태적 강제 할당 기법의 한계를 극복하기 위해 학계에서 획기적인 대안으로 제안된 것이 P&S (2009)의 합성품목번호(Synthetic ID) 기법이다. 이들은 개별 코드 간의 일대일 매칭이라는 기존의 강박에서 벗어나, 특정 분석 대상 기간 동안 분화되거나 수렴된 이력이 있는 모든 코드를 꼬리물기식으로 추적하여 하나의 거대한 가족 트리(Family Tree)로 묶어내는 발상의 전환을 시도하였다. 이들의 알고리즘은 미국 통계국(Census)이 매년 발간하는 신구 코드 대조표(new-obsolete files)를 입력 자료로 삼아, 한 시점의 코드가 여러 코드로 갈라져 나가는 'growing family tree' 형태와 여러 코드가 하나로 합쳐지는 'shrinking family tree' 형태, 그리고 이 둘이 결합된 형태까지 모두 추적한다. 이렇게 추적된 코드 가족에는 시점 정보가 결합된 고유한 합성 식별자, 즉 'setyear'[4]가 부여되며, 사용자는 임의의 시작 연도와 종료 연도를 지정하

4) P&S의 저자들이 STATA 코드에서 setyr로 명명한 변수를 의미. 본 논문에서는 이를 합성품목번호(syntheticID)라고 명명한다.

여 그 구간 내에서 연결되는 모든 코드를 단일한 합성 코드로 통합할 수 있다. 이 기법을 통해 연구자들은 인위적인 데이터의 소실이나 특정 연도의 비정상적인 교역액 급증 없이 15년 이상의 장기 시계열 데이터를 안정적으로 분석할 수 있게 되었으며, 해당 방법론은 미국 무역 데이터의 시계열 분석에 있어 사실상의 표준으로 자리 잡았다. Bernard, Jensen, Redding and Schott (2009)은 P&S (2009)의 연계표를 직접 활용하여 1993~2003년 미국 무역의 외연적·집약적 한계를 분해 분석한 첫 사례이며, 이후 P&S의 알고리즘은 동일 저자의 후속 SIC/NAICS 연계 연구인 P&S (2012)[5]와 결합되어 미국 무역·산업 데이터의 미시적 연결을 가능하게 하였다.

이 알고리즘의 적용 범위는 곧 미국 외 자료로 확장되었다. Van Beveren, Bernard, and Vandenbussche (2012)는 P&S의 알고리즘을 차용하여 EU의 8단위 결합명명(Combined Nomenclature, CN) 분류와 EU 산업생산통계(Prodcom) 분류에 동일한 방식의 시계열 합성 연계 절차를 구축하였으며, 일관된 합성품목번호를 사용할 경우 동일 기업의 제품 추가·퇴출이 과대측정되는 문제가 크게 완화됨을 벨기에 기업 자료로 입증하였다. 한편 Bellert and Fauceglia (2019)는 동일한 family tree 논리를 객체 지향(Java) 알고리즘으로 재구현하여 스위스 자료에 적용함으로써 코드 오류 가능성을 줄이는 실용적 방법을 제안하였고, 가장 최근의 Baumgartner, Srhoj, and Walde (2023)는 EU의 결합명명(CN)의 1995~2022 기간의 자료와 Prodcom의 2001~2021기간의 자료에 대한 합성 연계를 R 패키지[6]로 공개하여 후속 연구자가 즉시 활용할 수 있도록 하였다. 이러한 후속 연구는 P&S의 family tree 방식이 미국 자료에만 국한되는 기법이 아니라 국가·지역별 세분류 체계로 일반화될 수 있는 알고리즘이라는 점을 거듭 확인시켜 준다.

# III. 대한민국 HSK를 사용한 합성품목번호 시계열 연계

본 장은 2010년부터 2026년까지 17년간의 대한민국 HSK 10단위 코드의 모든 변동 이력을 추적하여 단일한 합성품목번호 연계표를 구축하는 절차를 단계적으로 서술한다. 본 연구는 P&S (2009)가 제시한 가족 트리(family tree) 식별 알고리즘의 수학적 본질을 그대로 유지하되, 한국 HSK 자료의 구조적 특성에 맞게 재구현한 형태로 적용하였다. 동일한 산출물을 두 가지 독립적인 알고리즘인 Python과 STATA로 각각 계산하여 행 단위로 교차 검증함으로써, 결과의 알고리즘적 정합성도 함께 확보하였다. Python과 STATA 코드와 그것들의 상세한 설명은 https://github.com/jayjeo/PS2009 에서 제공한다. 또한 최종적으로 구축한 합성품목번호 산출물을 시각적으로 편리하게 열람가능하도록 상기된 링크에서 exe 프로그램으로 제공한다. 다만, 상

5) P&S (2009)는 시간에 따라 분화·수렴되는 HS 코드를 가족 트리(family tree)로 추적하여 합성품목번호를 부여하는 HS-HS 시계열 연계 알고리즘을 다룬다. 반면, 동일 저자들의 P&S (2012)는 미국 HS 10단위 코드와 산업분류(SIC/NAICS) 사이의 매핑을 다루는 HS-산업분류 간 연계 논문이다.

6) 패키지명: harmonizer

기한 Python과 STATA 코드의 실행에 필수적인 “연도-연도간의 HSK 연계표”는 저자가 소속된 기관의 대외비 자료이므로 공개하지 못하였다.

## 1. 활용 데이터

본 연구의 입력 자료는 두 종류이다. 첫째는 분석 기간 17년간의 HSK 10단위 코드의 모든 신설·폐지·변경 이력을 통합한 신구 대조표 (즉, 연도-연도간의 HSK 연계표)이며, 둘째는 분석 기간 각 연도에 실제로 유효하였던 HSK 10단위 코드의 목록이다. 전자는 어떤 코드들이 같은 가족에 속하는지를 결정하는 자료이며, 후자는 각 연도에 그 코드가 실제로 법률상 존재하였는지를 결정하는 자료이다. 두 자료를 함께 사용하는 이유는, 신구 대조표만으로는 변경되지 않은 코드의 존재 여부를 확정할 수 없고 연도별 코드 목록만으로는 변경된 코드들의 가족 관계를 추적할 수 없기 때문이다.

신구 대조표의 각 행은 (변경 전 코드, 변경 전 시점, 변경 후 코드, 변경 후 시점)의 네 변수로 구성되며, 분류 체계상 변경 없이 동일 코드로 연속된 품목은 신구 대조표에 행이 존재하지 않는다. 다만 후술하는 바와 같이 이 원자료는 다수의 체계적 오류를 포함하고 있어, 본 연구는 알고리즘 적용을 하기 전에 먼저 제3장 2절에 설명된 정정 절차를 거쳤다. 정정 후 최종 입력으로 사용된 신구 대조표는 8,876행 규모이다.

연도별 유효 코드 목록 또한 단일 출처에 의존하지 않고 다중 출처 교차 검증 절차를 거쳐 구축하였다. 그 상세 절차는 역시 제3장 2절에서 함께 서술한다. 이렇게 정리된 연도별 코드 목록의 17년 합계 행 수는 202,682건이며, 매년 평균 약 11,922개의 HSK 10단위 코드가 유효하였다.

## 2. 원본 신구 대조표의 오류와 정정 절차

본 연구가 활용한 원본 신구 대조표는 그대로 합성품목번호 알고리즘의 입력으로 사용할 수 없는 두 종류의 체계적 오류를 포함하고 있음이 확인되었다. 이를 그대로 둔 채 알고리즘을 적용할 경우 가족 식별 결과가 부정확해지므로, 본 연구는 알고리즘 적용에 앞서 다중 출처 교차 검증과 양방향 정정 절차를 통해 신구 대조표를 정정하였다.

### 1) 원본 자료의 오류 유형

첫째 유형은 존재 시점의 비정합성 오류이다. 즉, 특정 연도의 관세율표에 실제로는 존재하지 않았던 HSK 코드가 마치 그 연도에 존속하는 것처럼 신구 대조표에 자기 매핑(self-loop, 변경

전 코드와 변경 후 코드가 동일한 행)으로 기재되어 있는 사례가 다수 발견되었다. 반대로 특정 연도의 관세율표에 실제로 존재하였던 HSK 코드가 신구 대조표에서는 그 시점에 이미 폐지된 것으로 잘못 기재되어 자기 매핑이 누락된 사례도 함께 발견되었다. 이러한 오류는 실무자가 직접 입력·교정하는 과정에서 발생하는 사무적 누락 또는 중복 기재에 기인한다. 이 오류가 알고리즘 결과에 미치는 영향은 결정적이다. 자기 매핑은 그 코드가 다른 매핑을 통해 같은 가족에 자동 포함되는지의 판단 근거가 되는데, 자기 매핑의 존재 여부 자체가 잘못 기록되어 있다면 그 코드가 실제로 어느 가족에 속해야 하는지에 대한 판정이 오염되기 때문이다.

둘째 유형은 형식적 오기 오류이다. 첫 자리가 0 (영)으로 시작하는 HSK 코드(주로 03류 수산물)에서 선행 영이 누락되어 9자리로 기록된 사례가 77건, 변경 후 코드가 빈 문자열로 기록되어 어떠한 후속 코드도 없는 것처럼 보이는 사례가 12건 발견되었다. 전자는 단순 형식 손상으로 의미상 변경이 아니나 그대로 두면 가족 식별 시 변경 후 코드가 다른 코드들과 매칭되지 않는 결과를 낳고, 후자는 그 변경 전 코드가 어떠한 후속 코드 없이 폐지된 것으로 잘못 인식되는 결과를 낳는다.

### 2) 수정 절차

위 두 유형의 오류를 모두 해소하기 위해 본 연구는 다음 두 단계의 정정 절차를 적용하였다.

**제1단계:** 신뢰 가능한 연도별 HSK 코드 목록의 독립적 구축을 먼저 하였다. 본 연구는 2010년부터 2026년까지 각 연도에 실제로 유효하였던 HSK 10단위 코드의 목록을 다음 네 가지 출처에서 독립적으로 수집하여 교차 검증하였다. 한 가지 출처만 사용하지 않았던 이유는 네 가지 출처끼리 상호비교를 한 결과, 각 출처들에서 모두 오류가 발견되었기 때문이다.

첫째, 법제처 국가법령정보센터에서 매년 공포되는 관세법 시행령 별표 및 관세율표 본문을 직접 다운로드하여 HSK 코드를 추출하였다. 둘째, 관세청 UNIPASS 웹페이지에서 HSK 코드 조회 결과를 Python을 사용하여 크롤링하여 별도의 연도별 코드 목록을 구축하였다. 셋째, CIEL(씨엘) 데이터베이스에서 연도별 HSK 코드를 Python을 사용하여 크롤링하였다. 넷째, 관세청 산하 CLIP(관세법령정보포털)에서 연도별 HSK 코드를  Python을 사용하여 크롤링하였다. 운영 편의상 2010년부터 2016년까지는 CLIP·법제처 자료를 1차 출처로, 2017년부터 2026년까지는 UNIPASS 자료를 1차 출처로 사용하였으며, CIEL 자료를 일관된 2차 출처로 활용하였다.

이렇게 수집한 1차 출처와 2차 출처를 비교대조하여 그 진위를 확인하였다. 분석결과 각 출처별로 다소 차이가 발견되었으며, 최종적인 선정은 경력 12년차 관세사의 분석을 통해 결정하였다. 1차 및 2차 출처 모두에서 존재하는 것이 옳은 HSK인 경우도 있었고, 1차 출처에는 존재하나 2차 출처에는 존재하지 않는 HSK인 경우에 1차 출처의 존재가 옳은 HSK인 경우도 있었다. 그 반대의 경우도 있었다. 그 결과로 본 연구가 신뢰 기준으로 사용한 연도별 HSK 코드 목록이 산출되었으며, 이 목록의 17년 동안의 합계 행 수가 202,682건이다.

**제2단계:** 양방향 검증을 통한 신구 대조표 정정을 한다. 위 1단계에서 구축한 연도별 HSK 신뢰 코드 목록을 기준으로, 원본 신구 대조표의 오기를 다음 두 방향에서 검증하여 정정하였다. 이 때 "자기 매핑(self loop)"이란, 연도-연도간 HSK 연계표에서, 동일 품번이 한 행에 존재하는 경우를 의미한다. 가령 한 행에서 품목A (2012) → 품목A (2015)로, 품목A가 스스로 매핑되는 경우이다.

순방향 검증은 각 자기 매핑 행의 코드가 변경 후 시점(위의 예에서 2015년)의 "연도별 HSK 신뢰 코드 목록"에 실제로 존재하는지를 점검한다. 자기 매핑 행이 존재하지만 그 코드가 해당 연도의 신뢰 목록에 등장하지 않는 경우, 이는 실제로 폐지된 코드를 마치 존속하는 것처럼 잘못 기재한 사례이므로 해당 행을 제거한다. 반대로 자기 매핑 행이 존재하지 않지만 변경 후 시점(위의 예에서 2015년)의 신뢰 목록에 여전히 존재하는 경우, 이는 실제로 존속하는 코드의 자기 매핑이 누락된 사례이므로 자기 매핑 행을 새로 추가한다.

역방향 검증은 같은 논리를 변경 전 시점(위의 예에서 2012년)과 그 시점의 "연도별 HSK 신뢰 코드 목록"을 기준으로 반복한다. 두 방향의 검증을 모두 통과한 행만을 최종 신구 대조표에 포함시켰다.

이상의 절차를 통해 산출된 최종 신구 대조표는 8,876행 규모이다. 종합하면, 가장 뒷자리가 0이 빠져서 10자리가 아니라 9자리인 경우는 8건이었는데 전부 수정하였으며, 가장 앞자리가 0이 빠져서 10자리가 아니라 9자리인 경우는 77건이었는데 전부 수정하였다. 자기 매핑 행의 오류 정정 결과, 다음 <표 Ⅲ-1>과 같은 비율로 오류가 존재하였으며, 모두 수정완료하였다.

〈표 Ⅲ-1〉 자기 매핑 행의 오류 비율

| 단위: % | 자기 매핑이 비존재 하는 것이 옳은 경우 | 자기 매핑이 존재 하는 것이 옳은 경우 | 합계 |
|---|---|---|---|
| 원본 연계표에서 자기 매핑이 비존재 | 73 | 6.95 | 79.95 |
| 원본 연계표에서 자기 매핑이 존재 | 0.02 | 20.03 | 20.05 |
| 합계 | 73.02 | 26.98 | 100 |

이러한 다중 출처 교차 검증과 양방향 정정 절차를 통해, 본 연구는 합성품목번호 알고리즘이 신뢰할 수 없는 입력 자료에 의해 왜곡된 가족 관계를 식별할 위험을 사전에 차단하였다. 본 정정 절차 자체도 본 연구의 방법론적 기여 중 하나에 해당한다.

## 3. 알고리즘 입력 단계의 자료 정제

위 절의 정정 절차를 거친 신구 대조표(8,876행)에 대해, 합성품목번호 알고리즘은 다음의 추가 정제 단계를 적용한다. 이 단계는 그래프 이론적 관점에서 가족 식별의 정확성을 위해 필요한 표준 절차이며, 이전 절의 정정과 달리 자료의 의미를 변경하지 않는다.

알고리즘은 변경 전 코드와 변경 후 코드가 동일한 1,479건의 자기 매핑 행을 가족 식별 단계에서 제거한다. 이러한 자기 매핑은 분류 체계 개정 시 한 코드가 그대로 존속하면서 동시에 그 코드의 일부가 새로운 별개 코드로 분화되어 신설되는 경우 발생한다. 예를 들어 5506900000 코드는 2014년에서 2017년으로 넘어갈 때 자기 자신은 그대로 존속하면서 5506400000이라는 새 코드가 함께 신설되었다. 이 경우 신구 대조표에는 “5506900000 → 5506900000”이라는 자기 매핑 행과 “5506900000 → 5506400000”이라는 일반 변경 행이 모두 기록된다. 자기 매핑은 가족 관계를 식별하는 그래프 정보를 추가하지 않으므로 본 알고리즘에서는 제거하되, 해당 코드가 일반 변경 매핑을 통해 같은 가족에 자동으로 포함되는지는 별도 단계에서 확인한다.

이러한 자기 매핑 제거를 거친 결과, 가족 식별에 실제로 사용되는 유효 변경 매핑은 7,397건(8,876 − 1,479)으로 정리되었다.

## 4. 변경 관계의 그래프 표현과 코드 가족의 식별

본 연구는 정제된 7,397건의 유효 변경 매핑을 무방향 그래프(undirected graph)로 표현한다. 그래프의 각 매핑은 한 간선(edge)에 해당하며, 매핑의 양쪽에 등장하는 두 HSK 코드는 그 간선의 두 끝점(node)에 해당한다. 매핑이 본래 갖는 방향성, 즉 변경 전 코드에서 변경 후 코드로의 방향은 가족 관계 식별 목적상 무관하므로 무방향으로 처리한다. 두 코드가 한 매핑에 함께 등장하였다는 사실 자체가 두 코드를 같은 가족으로 묶기에 충분한 근거가 되기 때문이다. 본 연구는 각 매핑에 숫자 1부터 시작하는 고유한 일련번호를 부여하였다. 이 일련번호는 다음 단계의 가족 식별자 초기값으로 사용된다.

본 단계의 핵심 과제는 위에서 구축한 그래프의 모든 연결 요소(connected component)를 식별하는 작업이다. 두 코드가 직접적인 한 매핑으로 연결되어 있거나, 혹은 중간의 다른 코드들을 매개로 한 일련의 매핑들에 의해 간접적으로 연결되어 있다면, 두 코드는 같은 연결 요소, 즉 같은 가족에 속한다.

이러한 직간접 연결의 의미를 다음의 구체적 예시로 설명한다. 아래 5건의 변경 매핑이 존재한다고 가정하자.

2011년: 2833299000 → 2833292010 [매핑 1]
2011년: 2833299000 → 2833292090 [매핑 2]
2014년: 2833292010 → 2833292011 [매핑 3]
2014년: 2833292010 → 2833292019 [매핑 4]
2022년: 2833292011 → 2833292015 [매핑 5]

직접 연결만 보면, 매핑 1과 매핑 2는 동일한 변경 전 코드 2833299000을 공유하므로 서로 묶이고, 매핑 3과 매핑 4는 동일한 변경 전 코드 2833292010을 공유하므로 서로 묶인다. 그러나 2833292010은 매핑 1에서는 변경 후 코드로, 매핑 3에서는 변경 전 코드로 등장하므로, 매핑 1·2의 묶음과 매핑 3·4의 묶음이 간접적으로 연결된다. 마찬가지로 2833292011이 매핑 3과 매핑 5를 매개하여 두 묶음을 다시 연결한다. 결과적으로 6개 코드(2833299000, 2833292010, 2833292090, 2833292011, 2833292019, 2833292015)와 5건의 매핑 전체가 단 하나의 가족으로 묶여야 한다.

이러한 모든 직간접 연결을 추적하기 위해 본 연구는 다음과 같은 반복적 최솟값 전파(iterative minimum propagation) 알고리즘을 적용하였다. 먼저 각 매핑이 가진 일련번호를 그 매핑의 임시 가족 식별자로 설정한다. 이 시점에서는 7,397개의 매핑이 각자 자기 자신만으로 구성된 7,397개의 임시 가족을 갖는다. 그 다음, 더 이상 어떤 식별자도 변하지 않을 때까지 다음 두 단계를 반복한다.

**단계 A:** 동일한 HSK 코드가 등장하는 모든 매핑들을 모아, 그들이 가진 임시 가족 식별자 중 최솟값을 모두에게 동일하게 부여한다. 이는 한 코드를 매개로 직접 또는 간접 연결되는 매핑들을 같은 가족으로 묶는 작업이다.

**단계 B:** 한 매핑의 양쪽 끝점에 있는 두 코드에게, 그 매핑이 가진 임시 가족 식별자를 동일하게 부여한다. 이는 한 매핑으로 직접 연결된 두 코드를 같은 가족으로 묶는 작업이다.

이 두 단계가 한 사이클을 구성하며, 한 사이클이 끝날 때 어떠한 식별자도 변하지 않은 시점이 곧 모든 직간접 연결이 완전히 식별된 수렴 시점이다. 본 연구의 한국 자료에 적용한 결과 12 사이클 후에 수렴하였다. 가장 큰 가족이 590개의 코드를 포함하기 때문에 단순 1:1 매핑 위주의 데이터셋보다 더 많은 사이클이 필요하다. 위의 5건 매핑 예시에 직접 적용해 보면, 정확히 3 사이클 후에 6개 코드 전체의 임시 가족 식별자가 같은 값으로 통일되어 단 하나의 가족으로 올바르게 식별됨을 확인할 수 있다.

수렴 후, 같은 임시 가족 식별자를 공유하는 코드들이 곧 하나의 가족을 구성한다. 본 연구는 이 단계에서 한 코드가 두 개 이상의 가족에 속하는 경우가 발생하지 않는지 자동 검증을 수행하였다. 만일 그러한 경우가 단 한 건이라도 발견되면 알고리즘의 정합성에 문제가 있다는 뜻이므로 즉시 작업이 중단되도록 설계되었다.

## 5. 합성품목번호의 여부와 최종 연계표 구성

가족 식별이 완료된 후, 본 연구는 각 가족에게 숫자 1부터 시작하는 정수를 순차적으로 부여하여 합성품목번호로 사용한다. 가족 내 모든 코드는 그 가족의 합성품목번호를 공유한다. 한국 HSK 자료의 경우 분석 기간 동안 한 번이라도 변경된 적이 있는 코드들에 대하여 총 1,809개의 가족이 식별되었으며, 이들 가족에 속하는 고유 HSK 코드는 7,816개이다. 평균 가족 크기는 약 4.32개 코드, 중앙값은 3개이며, 가장 큰 가족은 590개 코드를 포함한다. 가장 큰 규모를 갖는 가족은 8542(반도체 집적회로)·9031·9030·9027(계측·검사기기)·8524·8517(통신·디스플레이) 등 전기·전자·계측기 류이다. 이들은 17년에 걸쳐 누적적으로 분화·수렴을 많이 했기 때문에 이렇게 큰 가족이 형성된 것이다. 두 번째로 큰 가족(245개 코드)은 3824·3002·2933·3907 등 화학·의약품 류에 분포한다. 16개 이상의 코드를 포함하는 거대 가족은 36개이다.

여기서 한 가지 추가 처리가 필요하다. 위에서 식별된 가족은 분석 기간 동안 신구 대조표에 한 번이라도 등장한 코드, 즉 변경 이력이 있는 코드만을 포함한다. 그러나 17년간 단 한 번도 변경되지 않은 코드 또한 다수 존재하며, 이러한 코드들은 신구 대조표에 등장하지 않으므로 위 알고리즘에서 어떠한 가족에도 자동으로 배정되지 않는다. 본 연구는 이러한 변경 없는 코드 각각에도 고유한 합성품목번호를 부여하였다. 구체적으로는 가족에 부여된 합성품목번호의 최댓값(1,809) 다음 번호인 1,810번부터 시작하여, 변경 없는 코드들을 사전 순으로 정렬한 후 9,315번까지 순차적으로 부여하였다. 그 결과 7,506개의 변경 없는 코드가 각각 자기 자신만의 고유 합성품목번호를 갖게 된다. 즉, 이들의 경우는 한 개의 합성품목번호 당 오직 한 개의 HSK만 존재한다.

이 처리의 의의는 다음과 같다. 후속 무역 데이터 분석에서 연구자가 합성품목번호를 기준으로 데이터를 집계할 때, 변경된 코드는 가족 단위로 자동 합산되어 시계열 단절이 제거되는 한편, 변경 없는 코드는 개별 단위로 고유한 합성품목번호로 그대로 보존되어 미시적 정보가 손실되지 않는다. 어떤 코드라도 빠짐없이 합성품목번호의 분석 단위가 부여되며, 동시에 시계열 단절을 일으키지 않는 일관된 집계가 가능해진다.

마지막으로, 위 단계에서 산출한 합성품목번호 부여 결과를 본 장 제2절 2항에서 수집한 연도별 유효 코드 목록과 결합하여 본 연구의 최종 산출물인 연계표를 구성하였다. 이 연계표의 각 행은 “HSK 10단위 코드, 연도, 합성품목번호”의 세 변수로 구성되며, 그 의미는 “이러한 HSK 코드는 이 연도에 분류 체계상 유효하였고, 동시에 이 합성품목번호가 가리키는 가족에 속한다”는 것이다. 최종 연계표의 규모는 202,682개 행으로, 이는 17개 연도에 걸쳐 평균 약 11,922개의 HSK 코드가 매년 유효하였음을 의미한다. 합성품목번호의 총 개수는 9,315개이며, 그중 1번부터 1,809번까지는 한 번이라도 변경된 적이 있는 가족, 1,810번부터 9,315번까지는 17년 내내 변경되지 않은 개별 코드에 대응한다.

<표 Ⅲ-2>는 2010~2026년 전 기간에 걸쳐서, 한 개의 합성품목번호에 들어있는 HSK의 개수의 빈도수를 나타낸다. 가령 17년 내내 한번도 변경되지 않았던 HSK는 제2열 (1개)에 해당한

다. 표에서 볼 수 있듯이 한두 개 이하의 HSK인 경우가 압도적인 대부분을 차지한다.

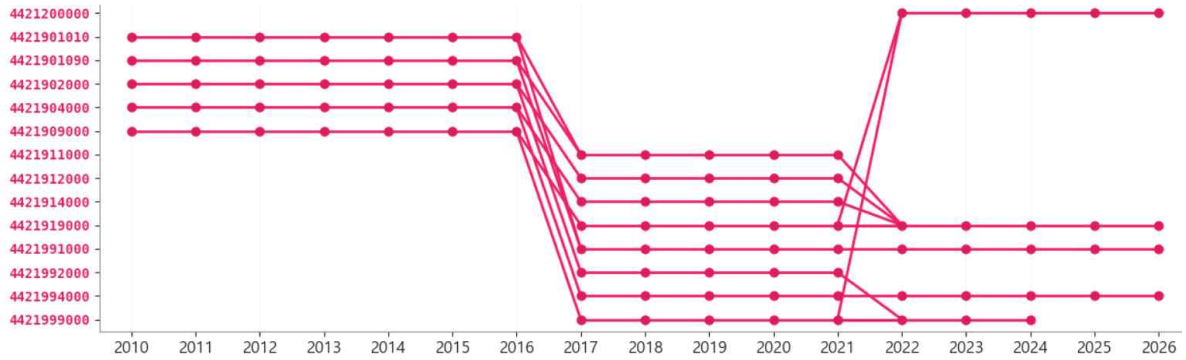
〈표 Ⅲ-2〉 한 개의 합성품목번호에 들어있는 HSK 개수의 빈도

| 한개의 합성품목번호안에 들어있는 HSK 개수 | 빈도수 |
|---|---|
| 1개 | 7,506 |
| 2개 | 766 |
| 3개 | 564 |
| 4개 | 156 |
| 5개~10개 | 262 |
| 11개~20개 | 35 |
| 21개~30개 | 11 |
| 31개 이상 | 15 |
| 합계 | 9,315 |

본 논문은 상기한 Github 링크에 합성품목번호 결과물 구축을 위한 Python 및 STATA 코드와 그것들의 상세한 주석서를 제공한다. 또한 합성품목번호 결과를 쉽게 열람할 수 있도록 exe 파일로 프로그램을 제공한다. 이하의 <그림 1, 2>는 동 프로그램 화면 일부의 스크린샷이다. <그림 1>에서 HSK = 3921199020의 합성품목번호는 4584번인데, 2010~2026 전 기간에 걸쳐서 유일한 HSK이다. 동 그림의 3921199020의 상단과 하단에 있는 6가지 HSK도 마찬가지다. 반면, 수렴, 분화하는 경우에는 한 개의 합성품목번호에 다수의 HSK가 존재하는 경우도 있다. 아래의 <그림 2>는 붉은 색으로 처리된 HSK 모두가 한 개의 합성품목번호 802번에 속한다.

〈그림 1〉 한 개의 합성품목번호 안에 한 개의 HSK만 존재하는 경우

〈그림 2〉 한 개의 합성품목번호 안에 다수의 HSK가 존재하는 경우




## 6. 산출 결과의 검증

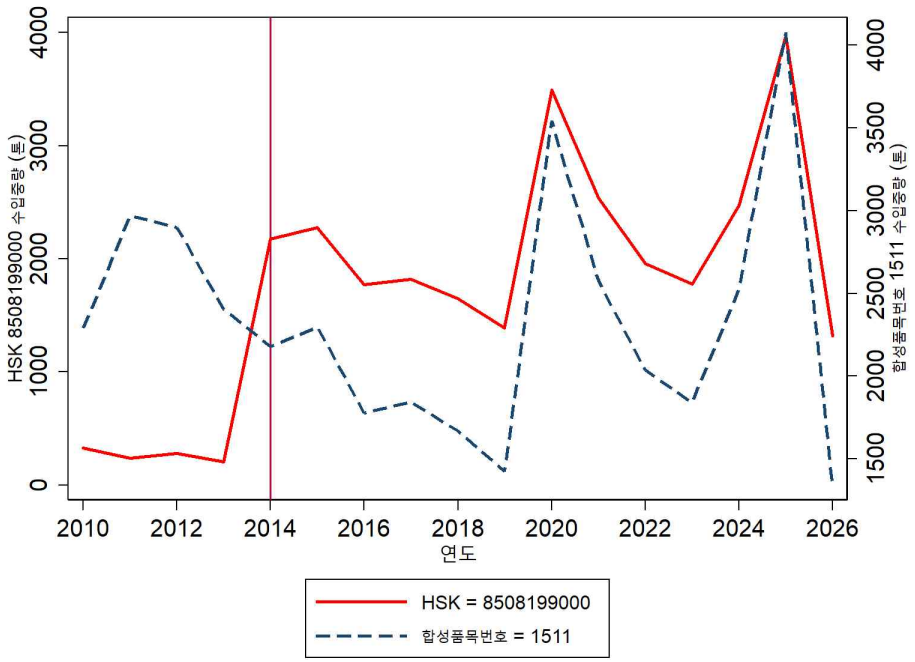
본 연구의 알고리즘은 알고리즘 내부 검증과 외부 독립 검증의 두 층위로 결과의 신뢰성을 확보하였다. 내부 검증으로는 세 가지 자동 점검을 알고리즘에 내장하였다. 첫째, 가족 식별 단계가 끝난 직후 한 HSK 코드가 두 개 이상의 가족에 속하지 않는지 확인한다. 둘째, 모든 유효 변경 매핑의 양쪽 끝 코드(변경 전 코드와 변경 후 코드)가 실제로 동일한 합성품목번호를 부여받았는지 확인한다. 한 매핑의 양쪽 코드가 서로 다른 합성품목번호를 갖는 상황은 가족 식별 알고리즘이 그 매핑이 의미하는 직접 연결을 누락했다는 뜻이므로 즉시 프로세스가 중단된다. 셋째, 최종 연계표의 모든 행에 합성품목번호가 결측 없이 존재하는지 확인한다. 한국 HSK 자료에 본 알고리즘을 적용한 결과 세 가지 검증 모두 통과하였다.

외부 독립 검증으로는 동일한 알고리즘을 완전히 다른 환경과 자료 구조에서 재구현하여 두 결과를 행 단위로 비교하였다. 본문에서 사용한 주 구현은 통계 패키지 STATA에서 위에 서술한 반복적 최솟값 전파 방식으로 작성된 것이다. 검증용 보조 구현은 Python의 networkx 라이브러리에서 제공하는 너비 우선 탐색(breadth-first search) 기반의 연결 요소 탐색 함수를 사용하였다. 두 알고리즘은 절차상 매우 다른 경로를 따르나, 무방향 그래프의 연결 요소를 식별한다는 동일한 수학적 문제를 푼다는 점에서 결과가 일치할 것으로 기대된다. 합성품목번호 부여 시점에서 두 구현이 모두 동일한 정렬 규칙, 즉, 각 가족이 포함하는 매핑 일련번호의 최솟값 기준을 채택하도록 설계하였으므로, 두 구현은 가족 분할뿐만 아니라 합성품목번호의 번호 자체도 동일한 값을 산출하도록 정렬된다. 두 구현의 결과 파일을 행 단위로 비교한 결과 모든 행에서 완전한 일치를 확인하였다. 이러한 다중 검증 절차를 통해 본 연구가 산출한 합성품목번호 연계표가 알고리즘적 오류 없이 구축되었음을 확인하였다.

## 7. 합성품목번호 기법의 실무적 유용성

본 절에서는 합성품목번호의 적용으로 얻을 수 있는 실무적인 이득에 대해서 예시를 들어 설명한다. 먼저 <그림 3>은 합성품목번호가 HSK의 수렴·분화로 인하여 발생하는 시계열 측정 착시를 어떻게 보정하는지를 한국 진공청소기(HS 8508)호의 실증 사례로 보여준다. 좌측 수직축의 빨간 실선은 단일 HSK 8508199000(기타 진공청소기)의 연도별 한국의 수입중량(톤) 시계열이다.[7] 2010년부터 2013년까지 이 코드의 수입중량은 약 200~330톤 수준에서 안정적으로 유지되었으나, 2014년에 갑자기 약 2,173톤으로 약 10.6배 폭증하는 양상이 관찰된다. 그러나 이 폭증은 실제 수요의 변화나 무역 정책 효과가 아니라, 2014년 관세율표 개정에서 형제 코드인 HSK 8508191000이 폐지되며 그 누적 무역량이 신설 코드 HSK 8508192000과 기존 코드 HSK 8508199000으로 분화 통합된 데 따른 측정 착시이다. 즉 연구자가 단일 HSK 8508199000만을 시계열 분석의 대상으로 삼는다면, 2013년과 2014년 사이의 이 인위적 단절을 실제 경제 현상으로 오인할 위험이 매우 크다.

〈그림 3〉 합성품목번호의 시계열 단절 보정 효과




7) 수입중량 자료는 관세청 수출입무역통계를 Python으로 크롤링하여 구축했다.

같은 그래프의 우측 수직축에 점선으로 표시된 합성품목번호 1511의 가족 합산 시계열은 이러한 단절을 완전히 해소한다. 합성품목번호 1511은 8508191000, 8508192000, 8508199000 세 개의 형제 코드를 단일 가족으로 묶어 관리하므로, 분류 체계 개정 시점에 한 코드의 무역량이 다른 코드로 이동하더라도 가족 합산 값은 그 영향을 받지 않는다. 실제로 가족 합산 수입중량은 2014년 분류 개정 시점(그래프상 붉은색 수직선)을 전후하여 약 2,400톤에서 약 2,180톤으로 사실상 변화 없이 매끄럽게 이어지며, 2010~2026년 17년 전체에 걸쳐 약 1,300~4,100톤 범위의 자연스러운 변동만을 보인다. 이로써 합성품목번호 기법이 단일 HSK 시계열에 내재된 분류 체계 개편으로 인한 측정 착시를 효과적으로 제거하고, 동일 상품군의 진정한 무역 흐름을 시계열 비교가 가능한 형태로 복원함을 본 연구의 한국 자료에서도 직접 확인할 수 있다.

한편, 합성품목코드는 HSK의 수렴 또는 분화로 인해서 연결된 이력이 있는 모든 HSK를 하나의 가족으로 묶기 때문에, 한 개의 합성품목코드 내에 있는 HSK들이 과연 동질성이 있다고 할 수 있는지 의구심을 제기 가능하다. 실무적으로 이러한 동질성은 중요한 경우가 많다. 대표적으로 덤핑방지관세의 부과를 예로 들 수 있다. 어떠한 특정한 품목(HSK)-국가에 덤핑방지관세를 부과하는 경우, 인접하고 유사한 HSK에도 동시에 덤핑방지관세를 부과하는 경우가 많다. 인접한 HSK에는 아무런 수입실적이 없었어도 예방 차원에서 그렇게 조치한다. 그렇게 하지 않으면 그러한 인접한 HSK로 우회덤핑이 쉽기 때문이다. 가령 품목A-중국에 덤핑방지관세를 부과할 때 동시에 품목B-중국, 품목C-중국에도 덤핑방지관세를 부과하는 경우, 기획재정부와 산업통상자원부의 무역위원회는 품목 A, B, C가 유사하다는 판단을 하고 있다고 볼 수 있다.

그런데 만약 어떤 합성품목번호가 있는데 그 안의 가족 HSK가 A, B, C, D, E라고 하자. 그리고 A, B, C는 앞서 예를 들었듯이 덤핑방지관세를 부과했는데, D, E에는 부과하지 않았다고 하자. 그런데 연구자가 HSK 개별단위가 아니라 합성품목번호 단위로 시계열 분석을 하고 있다고 하자. 이때 이 합성품목번호는 덤핑방지관세의 부과상태가 매우 애매하다. 부과되었다고 하기도 모호하고, 부과되지 않았다고 하기도 모호한 상태이다. 이런 상태가 빈번하게 발생한다면, 연구자는 합성품목번호를 사용하기 곤란하다.

이하의 <표 Ⅲ-3>은 각각의 “국가×연월×합성품목번호”에 대해서 덤핑방지관세가 부과중인 HSK의 비중을 나타낸다.[8] 가령 위의 예시에서 다섯 개의 HSK 중 A, B, C만 덤핑방지관세가 부과되고 있으므로 “덤핑방지관세를 부과중인 HSK의 비중”은 60%인 케이스이다. 따라서 이는 표에서 20<비중<80 에 해당하며, 빈도수는 1건이 추가된다. 표를 보면 비중이 0%인 경우의 건수가 64,422,150건으로, 전체의 99.96%를 차지함을 알 수 있다. 또한 비중이 100%인 경우의 건수가 24,550건으로, 전체의 0.04%를 차지한다. 이를 통해서 알 수 있는 것은 “애매한 경우”가 거의 존재하지 않는다는 사실이다. 즉, 합성품목번호 안에 들어있는 HSK 가족들은 거의 유사한 품목들로 이루어져 있음을 추측할 수 있다.

8) 국가는 한국향 수입실적이 존재하는 모든 국가, 연월은 2010년 1월에서 2026년 3월까지의 월별 자료이다.

〈표 Ⅲ-3〉 덤핑방지 관세를 부과중인 HSK의 비중

| 덤핑방지관세를 부과중인 HSK의 비중 (%) | 빈도수 (건) | 비중 (%) |
|---|---|---|
| 0 | 64,422,150 | 99.96 |
| 0< 비중≦20 | 77 | 0.00 |
| 20< 비중< 80 | 404 | 0.00 |
| 80≦비중< 100 | 31 | 0.00 |
| 100 | 24,550 | 0.04 |

# Ⅳ. Pierce and Schott (2009) 방법론의 구조적 한계점

미국 통계국(US Census)과 전미경제연구소(NBER)의 방대한 무역 자료를 바탕으로 고안된 P&S (2009)의 합성품목번호 기법은, 제품의 진입 및 퇴출에 관한 통계적 착시를 교정하고 장기 시계열 분석을 가능케 한 획기적인 공로를 지닌다. 그러나 이 방법론이 지닌 근본적인 논리, 즉 "교차 변경된 이력이 있는 모든 코드를 하나의 가족 트리(Family Tree)로 묶어 단일 코드로 취급한다"는 원칙은 분석 기간이 길어지고 적용 국가가 다양해질수록 극복하기 어려운 여러 가지 구조적, 통계적 오류를 야기한다. 본 장에서는 이 기법을 대한민국의 장기 HSK 데이터에 맹목적으로 적용할 경우 발생할 수 있는 주요 한계점들을 다섯 가지 차원에서 구체적으로 서술한다.

## 1. 과도한 데이터 집계로 인한 미시적 세분성(Granularity) 상실

P&S 방법론이 지닌 가장 치명적인 약점은 시계열이 길어질수록 단일 합성 코드의 규모가 비정상적으로 비대해지는 이른바 눈덩이 효과(Snowball effect)이다. 이 알고리즘은 1년 차에 A가 B와 C로 분화되고, 5년 차에 C가 D와 E로 분화되며, 10년 차에 E가 과거의 F와 합쳐지는 등의 일련의 과정이 존재할 경우, A부터 F까지의 모든 코드를 단 하나의 합성 코드로 묶어 버린다.

1~2년 정도의 짧은 기간에는 관련성이 높은 소수의 품목만 묶이게 되므로 산업의 특성을 유지할 수 있으나, 본 연구가 목표로 하는 2010년부터 2026년까지 무려 17년에 달하는 장기 시계열을 분석할 경우 사정은 달라진다. 기술의 융복합 현상과 WCO의 연쇄적인 분류 개정으로 인해 시간이 지날수록 전혀 다른 속성을 지닌 제품들이 우연한 교차점 하나 때문에 같은 가족 트리로 엮일 확률이 기하급수적으로 상승한다.

최신 연구인 Lukaszuk and Torun (2022)의 실증 분석에 따르면, P&S의 알고리즘을 사용하

여 1992년의 구형 HS 코드와 2017년의 최신 HS 코드를 연계할 경우, 가장 크게 형성된 단일 합성 그룹 하나에 무려 607개의 서로 다른 2017년 HS 코드가 통폐합되는 현상이 관찰되었다. 더욱 심각한 것은 이 거대한 607개의 코드 묶음 내부에 화학 물질(약 27%), 종이 제품(10%), 그리고 기계 및 전기 장비(40%) 등 산업적 연관성이 전혀 없는 완전히 이질적인 섹터의 품목들이 무차별적으로 뒤섞이게 된다는 점이다. 이러한 과도한 집계는 연구자가 개별 품목이나 세부 산업 단위에서 발생하는 미시적 경제 충격을 분석하고자 할 때, 합성 코드를 분석이 불가능한 블랙박스로 전락시키는 결정적인 원인이 된다.

하지만 제3장 7절의 예시에서 보듯이, 본 논문이 구현한 대상인 2010~2026의 HSK 연계에서, 덤핑방지관세 상태를 대리변수로 추정한 결과, 99.96%의 경우들이 동일 합성품목번호 내에서 모든 HSK들이 덤핑방지관세가 없는 상태, 거의 나머지인 0.04%의 경우들이 동일 합성품목번호 내에서 모든 HSK들이 덤핑방지관세를 받고 있는 상태였다. 이는 유사하지 않은 HSK가 동일한 합성품목번호로 들어오는 경우가 거의 존재하지 않는다는 의미로 간접적으로 해석할 수 있다.

다만, 위에서 지적한 한계를 근본적으로 극복하기 위해 가장 최근 스위스 장크트갈렌 대학 및 스위스 경제부(SECO) 소속의 Lukaszuk and Torun (2022)에 의해 가중치 기반 무역 흐름 배분 기법이 새롭게 제안되었다. 이들은 국가 간 교역되는 개별 품목의 무역 비중이 시간의 흐름에 따라 일정하게 유지되는 경향이 있다는 무역 데이터의 지속성에 착안하였다. 이들은 분화된 코드를 단순히 하나의 합성 카테고리로 묶어 버리는 대신, 분류 체계가 변경되기 전과 후의 양국 간 품목별 교역량 데이터의 제곱 편차를 최소화하는 제약조건부 최소제곱 최적화 알고리즘을 개발하였다. 이 최신 기법은 과거의 단일 코드에서 여러 코드로 쪼개져 나간 무역액을 추정된 가중치에 따라 정밀하게 분할 할당함으로써, P&S 기법이 훼손하였던 미시적 데이터의 세분성을 보존하면서도 시계열의 일관성을 유지하는 획기적인 성과를 보여주고 있다.

## 2. 통계적 분포의 극단적 왜곡 발생

품목들이 하나의 합성 코드로 병합되는 과정은 데이터의 횡단면적 분포를 심각하게 훼손한다. P&S 알고리즘의 특성상, 분석 기간 내내 단 한 번도 분류 기준이 변경되지 않은 안정적인 코드들은 기존의 독립적인 지위와 상대적으로 작은 교역 규모를 그대로 유지한다. 반면, 기술 변화가 빠르거나 세관의 행정적 수요에 의해 빈번한 쪼개짐과 합쳐짐을 겪은 품목들은 모두 하나의 거대한 합성 코드로 빨려 들어가면서 수십, 수백 개 품목의 교역액을 독식하게 된다.

그 결과, 전체 교역 데이터의 분포가 소수의 합성 코드에 극단적으로 편중되는 심각한 비대칭성(Skewness)이 발생한다. Lukaszuk and Torun (2022)에 의하면, 2018년 전 세계 교역 데이터에서 가장 거래 규모가 큰 품목 코드가 전체 교역액에서 차지하는 실제 비중은 본래 약 6.66% 수준에 불과했다. 그러나 P&S 방법론을 적용하여 HS 2017 코드를 HS 1992 코드 체계로 합성

그룹화하자, 이 가장 큰 묶음이 전체 교역액에서 차지하는 비중이 약 25.74%까지 인위적으로 폭등하는 결과가 초래되었다.

이러한 인위적인 통계적 거인의 탄생은 무역 경제학에서 흔히 사용되는 허핀달-허쉬만 지수(Herfindahl-Hirschman Index, HHI)나 산업 집중도(Concentration ratio)와 같은 시장 독점도 측정 지표들을 사실상 무용지물로 만든다. 연구자는 특정 산업의 소수 품목 집중도가 높아진 것이 시장의 실제 독점화 때문인지, 아니면 P&S 알고리즘이 빚어낸 코드 통폐합의 결과인지 구분할 수 없게 되기 때문이다. 따라서 합성품목분류를 기준으로 무역 데이터를 분석하는 경우 HHI 지수 등의 사용은 피하는 것이 좋다.

## 3. 표본 크기의 절대적 축소를 유발하는 레벨 효과(Level Effect)

무역 데이터의 미시적 분석에 있어 국가가 얼마나 다양한 종류의 품목을 수출입하는지를 나타내는 제품의 다양성(Product variety)은 후생 경제학적 관점에서 국가의 무역 경쟁력과 소비자 효용을 측정하는 핵심 지표이다. 그러나 P&S 기법은 필연적으로 다수의 개별 HS 코드들을 소수의 합성 코드로 병합해 버리기 때문에, 분석 데이터셋에 남아 있는 전체 품목 코드의 개수(Sample size)를 극적으로 축소시키는 부작용, 즉 '레벨 효과(Level effect)'를 동반한다.

아래 <표 Ⅳ-1>에서 나타나듯, P&S의 방법론은 개정 연도마다 발생하는 품목 수의 단발성 변동을 제어하는 데에는 성공하였으나, 반대급부로 분석 가능한 전체 품목의 범주를 무려 4분의 1 (25%) 가량 영구적으로 줄여 버리는 결과를 초래했다 (Lukaszuk and Torun, 2022). 이는 실제 글로벌 시장에서 기업들의 혁신과 무역 장벽 철폐로 인해 거래되는 신제품의 종류가 비약적으로 증가하여 실제 제품 다양성이 높아졌음에도 불구하고, 연구자의 통계 모형 상에서는 오히려 국가가 취급하는 제품의 종류가 영구적으로 감소한 것처럼 보이게 만드는 심각한 분석적 오류를 낳는다.

〈표 Ⅳ-1〉 각 연구자의 연계방법 및 통계적 왜곡

| 제안자 및 연계 방법 | 1992~2018년 기간 중 관측 가능한 품목 수의 변화 양상 | 통계적 왜곡의 결과 |
|---|---|---|
| UN Comtrade (단순 후방 변환) | 1992년에서 2018년까지 약 26년에 걸쳐 수입품 수가 약 10% 점진적으로 감소하며, 2007년 HS 개정 시 단발성 스파이크(자가신고 약 15%, Comtrade 변환 약 12%)도 동시에 발생 | 개정 시점마다 품목이 소멸한 것으로 착시 유발 |

| 제안자 및<br>연계 방법 | 1992~2018년 기간 중<br>관측 가능한 품목 수의 변화 양상 | 통계적 왜곡의 결과 |
|---|---|---|
| P&S<br>(합성품목코드) | 전체 시계열 구간에 걸쳐 관측 가능한 총 품목 수가 원본 데이터 대비 약 25% 가량 영구적으로 축소된 상태 유지 (Level effect) | 실제 시장의 제품 다양성을 심각하게 과소 추정 |
| Lukaszuk and Torun<br>(가중치 배분) | 분석 기간 전체에 걸쳐 품목 수의 인위적 하락이나 축소 없이 안정적 수준의 품목 다양성 유지 | 미시적 세분성과 시계열 일관성 동시 확보 |

## 4. 포괄적 매핑의 강제 삭제로 인한 신규 교역 흐름의 데이터 절단

이 알고리즘은 수학적 엄밀성을 유지하기 위해 코딩 과정에서 다루기 까다로운 예외적 데이터들을 과감하게 삭제(Drop)해 버리는 태생적 결함을 가지고 있다. P&S (2009)은 알고리즘을 설계할 때 미국 통계국 자료 중 뒷자리가 'X'로 표기되는 포괄적 매핑(Blanket mappings, 예: 8486XXXXXXX) 코드들을 분석 대상에서 일괄 삭제하도록 설계하였음을 명시적으로 밝히고 있다.

관세 행정의 실무적 관점에서 볼 때, 완전히 새로운 기술이 적용된 신제품(예: 초기 형태의 웨어러블 디바이스나 특수 신소재 등)이 시장에 처음 등장하여 아직 명확한 세부 10단위 분류 기준이 확립되지 않았을 때, 세관원들은 임시방편으로 이러한 X가 포함된 포괄적 코드를 사용하여 무역액을 집계하는 경향이 있다. 따라서 알고리즘의 원활한 구동을 위해 이러한 포괄적 매핑 관측치를 일괄 삭제한다는 것은, 무역 동향 분석에서 가장 중요하게 다루어져야 할 첨단 신제품들의 초기 진입(Entry) 흐름과 그로 인한 외연적 한계(Extensive margin)의 팽창 데이터를 통째로 날려 버리는 치명적인 데이터 손실(Data truncation)을 의미한다. 이 한계는 P&S (2009) 자신들도 결론부에서 명시적으로 인정한 바와 같이, 새로운 HS 코드는 언제나 기존 코드에서 분기되어 나오며 실제 상품이 시장에 등장한 후 한참 뒤에야 통계 당국이 별도 코드를 부여할 수 있다는 또 하나의 근본적 제약과 결합하여, 합성품목번호 알고리즘이 진정한 신상품(genuine new product)의 출현 시점을 정확히 포착하기에는 본질적으로 부적합하다는 사실을 시사한다.

다만 본 논문이 구현한 2010~2026년의 HSK의 합성품목번호 작업에서 살펴본 바에 따르면, 이하의 어떤 데이터에서도 한국 자료의 경우 임시방편으로 X와 같은 미정 상태가 존재하지 않았음을 밝힌다. 관세청 수출입 무역통계에서 가져온 국가별×연월별×HSK별 수출입데이터, 법제처 국가법령정보센터 관세율표 & 관세청 UNIPASS & CLIP(관세법령정보포털) & CIEL(씨엘)이 제공하는 각각의 모든 종류의 데이터에서 X와 같은 미정상태는 없었다.

## 5. 분석기간 의존성 및 개도국 데이터 적용 시의 행정적 시차 미반영 문제

P&S (2009)의 알고리즘은 사용자가 임의로 지정한 시작 연도와 종료 연도를 기준으로 그 구간 내에서 연결되는 가족 트리만을 식별하여 합성 코드(setyear)를 부여한다. 이는 동일한 코드 체계라 하더라도 분석기간을 2010~2026년으로 잡을 때와 2012~2022년으로 잡을 때 가족 트리의 범위와 합성 코드 구성이 달라질 수 있음을 뜻한다. 다시 말해 합성품목번호는 절대적이고 보편적인 상품 번호가 아니라, 특정 연구 윈도에 맞춰 생성되는 분석용 식별자라는 본질을 갖는다.

이러한 분석기간 의존성과 더불어 P&S 기법은 미국의 선진화된 관세 행정, 즉 WCO의 상위 6단위 개정 지침이 발표되면 당해 연도 1월 1일을 기점으로 즉각적으로 하위 10단위 코드가 일제히 갱신된다는 가정을 기저에 깔고 있다. 그러나 세계은행의 Cebeci (2012)가 정당하게 비판하였듯, 이 방법론은 행정력이 부족한 개발도상국의 무역 데이터를 연계할 때 심각한 오류를 발생시킨다. 상당수의 국가들은 새로운 HS 분류 체계가 도입되더라도 기존 관세 시스템을 즉각 업데이트하지 못하고, 구형 코드와 신형 코드를 수년 동안 혼용하여 기재하는 경우가 빈번하다.

P&S 알고리즘은 사전에 연구자가 정해 놓은 특정 시작 연도와 종료 연도(예: 1989~2004년)의 신구 대조표(new-obsolete files)를 기준으로 엄격하게 가족 트리를 형성하므로, 구형 코드가 공식 폐기 시점 이후에도 계속해서 통계에 등장하는 이러한 행정적 시차와 중첩 현상을 유연하게 처리하지 못한다. 그 결과 연결 고리가 끊어지거나 잘못 매핑되어 특정 국가의 교역액이 통째로 누락되는 현상이 발생하게 된다.

다행히 한국의 HSK는 관세청 고시에 의해 압도적 대부분의 경우 매년 1월 1일자로 일제히 갱신되며, 분석기간 17년을 통틀어 통상적이지 않은 월에 고시된 사례는 단 1건[9]에 불과하므로 이러한 행정적 시차 문제는 미국과 유사한 수준으로 통제 가능하다.

또한 한국의 경우 저자가 오랜 경험적으로 아는 한, Cebeci (2012)가 제기한 문제가 발생하는 경우는 거의 없다. 즉, 행정력이 부족한 개발도상국은 새로운 HS 분류 체계가 도입되더라도 기존 관세 시스템을 즉각 업데이트하지 못하고, 구형 코드와 신형 코드를 수년 동안 혼용하여 기재하는 문제가 있지만, 저자의 확인에 따르면 관세청 수출입무역통계는 이하에서 설명할 한 가지 문제만 제외하면 이러한 문제가 존재하지 않는다.

관세청 수출입무역통계를 자세히 뜯어보면, 특정연도에 HSK가 폐지되어 더 이상 존재하지 않아야만 하는 경우에도 수출입 실적이 잡히는 경우가 상당수 존재한다. 그런데 이것은 “구형 코드와 신형 코드를 수년 동안 혼용하여 기재”한다는 Cebeci (2012)가 지적한 개발도상국 형 문제가 아니다. 저자의 추측에 따르면, 관세청 수출입 실적에서 이러한 문제가 발생하는 이유는 신구형 코드의 혼용 때문이 아니라, 일부 수출입 신고가 몇 달 늦게 반영되는 행정적인 원인 때문이다. 가령 HSK의 품목A가 2015년까지만 존재했고 2016년에 폐지되었는데, 신고자가 2015년

9) 2019년에 예외적으로 1월 1일에 공포되었다가 또 다시 10월에 공포된 바 있다.

말에 품목A로 신고를 한 경우, 일부 신고가 늦게 처리되어 2016년 1월에 등장하는 경우이다. 이러한 노이즈(Noise) 데이터는 시계열의 구조적 정합성을 훼손하므로, 본 논문의 저자는 무역실적을 활용하는 다른 연구를 수행할 때 합성품목번호 연결 요소 알고리즘을 적용하기 전 단계에서 해당 이상치 관측치를 분석 대상에서 통제하였다. 다만 본 연구에서는 무역실적(수출입 통계)을 활용하지 않았으므로, 위의 문제와 아무 연관이 없다.

한편, 법제처 국가법령정보센터 관세율표 & 관세청 UNIPASS & CLIP(관세법령정보포털) & CIEL(씨엘)에서 제공하는 모든 정보를 각각 다운로드하여 저자가 서로의 데이터를 비교대조한 결과, 제공처가 수기로 업데이트 하는 과정에서 다소의 오류가 존재함을 확인했다. 예를 들면 CLIP 자료에는 2016년부터 품목A가 폐지된 것으로 나오는데 CIEL 자료에는 2016년에도 품목A가 여전히 존재하는 것으로 나타난다. 그런데 이것은 Cebeci (2012)가 지적한 신구형 코드의 혼용문제가 아니라, 단순히 각 자료제공 기관의 시스템적인 비동기화 문제일 뿐이다. 주의할 것은, 한국 수출입이나 HSK 목록 데이터를 얻을 때, 한 기관만을 믿고 사용하면 안되고, 여러 기관의 동일 자료를 중복으로 확보하여 비교대조한 후, 관세사의 판단 아래 무엇이 옳고 무엇이 잘못된 정보인지를 판단하는 작업을 반드시 거쳐야 한다.

## 6. 정책변수 부착 시 합성 그룹 내 이질성 (Within-group Heterogeneity)

합성품목번호로 묶인 그룹은 분류 체계상의 시계열 일관성을 보장할 뿐, 그 그룹에 부여되는 정책 변수의 동질성까지 자동으로 보장하지는 않는다. 분석 기간이 길어질수록 한 합성 코드 안에 서로 다른 관세율, 자유무역협정(FTA) 특혜 적용 여부, 수입 규제 대상 여부, 원산지 규정, 전략물자 통제 여부 등이 혼재될 가능성이 커지며, 이는 곧 정책 처치(policy treatment) 변수를 합성 그룹 단위로 단순 부착하기 어려운 상황을 초래한다.

앞서 제2장 2절에서 살펴본 KIEP의 정재욱·김예진 (2018) 보고서는 바로 이 문제에 대한 모범적 대응 사례를 보여준다. 동 보고서에서는 미국 HS 6단위 합성 그룹 내부에 AGOA 수혜 품목과 비수혜 품목이 혼재되는 현상이 관찰되었고, 연구진은 이를 0과 1의 단일 더미 변수로 부착할 수 없다고 판단하여 그룹 내부의 미국 대(對)세계 수입 비중으로 가중한 0과 1 사이의 연속형 수혜율 변수를 별도로 설계함으로써 정책 처치의 강도를 정량적으로 표현하였다. 즉 합성 코드 자체가 코드 연계 문제를 해결하더라도 정책 변수의 깔끔한 부착 문제까지 자동으로 해결해 주지는 못한다는 점이 실증적으로 입증된 셈이다. 본 연구가 향후 한국 HSK 합성품목번호 연계표를 작성하고 이를 FTA 효과 분석, 관세율 정책 평가, 수입 규제 효과 추정 등에 활용하고자 한다면, 합성 코드와 별개로 각 정책 변수의 그룹 내 이질성 정도를 측정하는 보조 가중치 체계를 함께 설계할 필요가 있음을 KIEP의 선례가 명확히 시사하고 있다.

단, 본 논문이 제3장 7절에서 보인 바와 같이, 각 합성품목번호 내에서 덤핑방지관세가 부과되는 HSK와 부과되지 않는 HSK가 혼재하는 경우는 거의 존재하지 않는다. 따라서 사용하고자

하는 효과분석의 대상에 따라 연속형 변수로 가중치를 사용하든지, 이산형 변수를 유지하든지 연구자가 적절히 선택하면 될 것이다.

# V. 결론

국제 무역의 실증적 시계열 분석에서 HSK 분류 체계의 잦은 분화와 수렴으로 인해 발생하는 인위적 데이터 값의 단절을 해결하는 작업은 분석 결과의 신뢰성을 담보하기 위한 절대적 선결 과제이다. 그러나 공개적으로 확인 가능한 범위 내에서 P&S (2009)의 합성품목번호 기법을 한국 HSK 10단위에 직접 적용하여 시계열 연계표 자체를 학술적 산출물로 제시한 국내 연구는 부재한 상태였다. 기존 국내 KCI 및 국제 SSCI급 학술지 논문은 미국 등 타국의 연계 데이터를 이차적으로 활용하거나 광범위한 산업 분류로 통계를 집계한 거시적 접근에 머물렀으며, KIEP의 정재욱·김예진 (2018)과 같은 일부 국책연구기관 보고서에서조차 미국 HS 6단위 수준에서 제한적으로 동 기법을 원용하는 데 그쳤다. 본 연구는 이러한 학술적 공백을 메우기 위해 2010년부터 2026년까지 17년간의 한국 HSK 10단위 전 품목을 대상으로 P&S의 합성품목번호 알고리즘을 직접 적용하여 단일하고 일관된 시계열 연계표를 구축하였다.

본 연구의 방법론적 기여는 세 가지로 요약된다. 첫째, 원본 신구 대조표가 갖는 사무적 오류(자기 매핑의 누락 또는 잘못 기재, 형식적 오기)를 해소하기 위해 법제처 국가법령정보센터, 관세청 UNIPASS, CIEL, CLIP의 네 가지 독립 출처를 교차 검증하고, 이를 토대로 양방향(순방향·역방향) 검증을 통한 신구 대조표의 정정 절차를 자체적으로 설계하여 적용하였다. 그 결과 정정된 신구 대조표 8,876행을 입력으로 하여 1,809개의 가족과 7,506개의 변경 없는 개별 코드, 총 9,315개의 합성품목번호와 202,682행의 최종 연계표를 산출하였다. 둘째, 동일한 알고리즘을 STATA (반복적 최솟값 전파 방식)와 Python (networkx의 너비 우선 탐색 방식)의 두 가지 독립 구현으로 작성하고 두 결과를 행 단위로 비교하여 완전한 일치를 확인함으로써 산출물의 알고리즘적 정합성을 보장하였다. 셋째, 양 구현의 코드와 그 상세 주석서, 그리고 합성품목번호 결과를 시각적으로 조회할 수 있는 실행 파일(exe) 형태의 프로그램을 GitHub에 공개하여 후속 연구자들이 즉시 활용할 수 있는 기초 데이터 인프라로서의 접근성을 확보하였다.

본 연구가 산출한 합성품목번호 연계표의 실무적 유효성은 두 가지 실증 사례를 통해서도 확인된다. 진공청소기(HSK 8508199000)의 사례에서 단일 HSK 시계열만으로는 2014년에 약 10.6배의 측정 착시가 발생하지만, 합성품목번호 가족 합산 시계열에서는 그러한 단절 없이 매끄럽게 이어진다. 또한 덤핑방지관세 부과 상태를 대리변수로 하여 합성품목번호 내 HSK들의 정책적 동질성을 검정한 결과, 전체 약 6,447만 건의 “국가×연월×합성품목번호” 단위 관측치 중 사실상 거의 모두에서 가족 내 모든 HSK가 동일한 부과 상태 (99.96%는 모두 미부과, 0.04%는

모두 부과)를 공유하는 것으로 나타나, 합성품목번호로 묶인 가족이 정책적으로도 매우 동질적인 묶음임이 간접적으로 확인되었다. 아울러 한국의 HSK는 매년 1월 1일자 일괄 개정[10]으로 미국과 유사한 수준의 행정적 정시성을 갖추고 있어, Cebeci (2012)가 개발도상국 자료에 대해 지적한 신구 코드 혼용 문제로부터도 사실상 자유로움을 함께 확인하였다.

본 연구가 제공하는 한국 HSK 합성품목번호 연계표는 향후 한국 무역 데이터의 미시적 시계열 분석을 시도하는 후속 연구자들에게 시계열 단절 문제로부터 자유로운 견고한 기초 데이터셋의 역할을 수행할 것으로 기대된다. 다만 P&S의 합성품목번호 기법은 분석 기간이 매우 길어질수록 단일 가족의 규모가 비대해질 가능성을 일반적 한계로 내포하고 있으며, 이를 가중치 기반 무역 흐름 배분 방식으로 보완한 Lukaszuk and Torun (2022)의 최신 연구가 별도로 제안되어 있다. 본 연구의 한국 자료에서는 해당 한계가 실증적으로 미미한 수준으로 확인되었으나, 보다 정밀한 미시 가중 분석이 필요한 후속 연구에서는 이러한 대안적 기법을 함께 검토할 수 있다. 즉, 본 논문에서 구축한 1차적 HSK 연계표 인프라를 바탕으로, 향후 연구에서는 가중치 기반 무역 흐름 배분 모형을 결합하여 합성 그룹 내 무역액을 정밀하게 분할하는 2차적 연계표 모형으로 진화할 수 있을 것이다.

10) 단, 2019년 만은 예외적으로 1월 1일에 공포되었다가 또 다시 10월에 공포된 바 있다.

# 참 고 문 헌

김중선·심상렬 (2022). 한국 반도체 산업의 표준품목분류 체계 연구: MTI와 HS 코드 비교 분석을 중심으로. **관세학회지**, 23(2), 79 - 99.

김진규 (2022). 제7차 HS 협약 개정에 따른 HSK 품목분류 신설에 관한 연구: HS 제3907호를 중심으로. **관세학회지**, 23(1), 25 - 44.

박윤수·전태완·신선경·엄남일 (2022). 수출·입 폐기물의 관리 개선 마련 연구: 관세·통계통합품목분류표(HSK)를 중심으로. **한국폐기물자원순환학회지**, 39(6), 489 - 497.

정재욱·김예진 (2018). 미국 아프리카성장기회법(AGOA)의 교역 효과와 정책적 시사점 (전략지역심층연구 18-01). **대외경제정책연구원(KIEP)**.

Baumgartner, C., Srhoj, S., & Walde, J. (2023). Harmonization of product classifications: a consistent time series of economic trade activities. Jahrbücher Für Nationalökonomie Und Statistik, 243(6), 643 - 662.

Bellert, N., & Fauceglia, D. (2019). A practical routine to harmonize product classifications over time. International Economics, 160, 84 - 89.

Bernard, A. B., Jensen, J. B., Redding, S. J., & Schott, P. K. (2009). The margins of US trade. American Economic Review, 99(2), 487 - 493.

Cebeci, T. (2012). A concordance among Harmonized System 1996, 2002 and 2007 classifications (World Bank Working Paper No. 74576). World Bank.

Feenstra, R. C. (1996). U.S. imports, 1972-1994: data and concordances (NBER Working Paper No. 5515). National Bureau Of Economic Research.

Frazer, G., & Van Biesebroeck, J. (2010). Trade growth under the African Growth and Opportunity Act. The Review Of Economics And Statistics, 92(1), 128 - 144.

Lukaszuk, P., & Torun, D. (2022). Harmonizing the Harmonized System (SSRN Working Paper No. 4302540). University Of St. Gallen / SECO.

Pierce, J. R., & Schott, P. K. (2009). Concording U.S. Harmonized System categories over time (NBER Working Paper No. 14837). National Bureau Of Economic Research.

Pierce, J. R., & Schott, P. K. (2012). A concordance between ten-digit U.S. Harmonized System codes and SIC/NAICS product classes and industries. Journal Of Economic And Social Measurement, 37(1-2), 61 - 96.

Van Beveren, I., Bernard, A. B., & Vandenbussche, H. (2012). Concording EU trade and production data over time (NBER Working Paper No. 18604). National Bureau Of Economic Research.

## ABSTRACT

# Time-Series Linkage of Korea's HS codes Using Synthetic ID

**Deokjae Jeong**

The analysis of trade statistics suffers from time-series discontinuities arising from the frequent divergence and convergence of Harmonized System (HS) codes. To overcome this limitation, Pierce and Schott (2009) proposed the synthetic product code methodology as a groundbreaking alternative. However, to the best of the author's knowledge, no domestic or international paper or report has applied their methodology to the HSK (Harmonized System of Korea). To fill this gap, this paper constructs a consistent time-series concordance table by applying the synthetic product code methodology of Pierce and Schott (2009) to the entire universe of South Korea's 10-digit HSK product codes over the 17-year period from 2010 to 2026. The implementation code is publicly released on GitHub[11] in both Python and Stata versions. To enable broad accessibility, the resulting concordance is further packaged as an executable (.exe) application that offers interactive visual querying and is publicly released on the same GitHub repository, providing subsequent researchers with a reference resource for micro-level time-series analyses of Korean trade data.

JEL Code: C81, F14, F13, C55, F10

Key Words : Synthetic Product Code, Harmonized System of Korea (HSK), Time-series Concordance, Pierce-Schott Algorithm

11) https://github.com/jayjeo/PS2009