스캐너 데이터 기반 물가 변동 분석

김용대, 이영환, 백규승, 김성현

서울대학교 통계학과/ 한국은행

연구주제

- 빅데이터를 기반으로 물가지수를 실시간으로 모니터링 하려는 시도가 진행 되고 있음.
- 개별 마트와 슈퍼의 스캐너 데이트를 분석하면 빠르게 물가지수를 구할 수 있음.
- 현재 이용 가능한 (구매 가능한) 스캐너 데이터의 문제점
  - 결측치가 많음
  - 데이터의 용량이 커서 일반 PC에서 분석이 어려움
- 본 연구의 목적: 행렬분해를 이용한 스캐너 데이터 결측치 보정

스캐너 데이터

- **스캐너 데이터**: 2013년 1월부터 2019년 9월까지 6년 9개월 (352주간) 3,241개의 도소매점에서의 각종 상품의 주별 판매량과 판매액 등의 판매 정보를 포함
- 약 285,000개의 개별 상품의 각 상점에서의 주별 판매 정보 로그 8억 4천만건으로 구성
- 로그: *상품군*, *상품명*, *판매 상점*, *판매량*, *판매금액* 으로 구성
- 목표: 스캐너 데이터를 이용해 물가 변동 추이를 분석

```
WEEK_END_DT,STORE_CD,BARCODE,KAN,KAN_NM,ITEM_NM,SUM_QTY,SUM_AMT
"20131208"|"2030BC"|"8015352307303"|"010101"|"식초"|"코리아제니스 피니발사믹글레이즈215ML - PET BOTTLE"|1|11700
"20131208"|"2030BC"|"8015352309956"|"010101"|"식초"|"코리아제니스 피니발사믹식초원립퀄리티500ML - BOTTLE"|1|5400
"20131208"|"2030BC"|"8801005213042"|"010101"|"식초"|"샘표 백년동안 산수유 석류 흑초 49.5% PLT.BTL 900ML"|1|9800
"20131208"|"2030BC"|"8801005214018"|"010101"|"식초"|"샘표식품 백년동안1/2산머루복분자흑초900ML - PLT BOTTLE"|1|11600
"20131208"|"2030BC"|"8801005214025"|"010101"|"식초"|"샘표식품 백년동안1/2산수유석류흑초900ML - PLT BOTTLE"|1|11600
"20131208"|"2030BC"|"8801007168739"|"010101"|"식초"|"씨제이제일제당 백설국내산사과로만든2배사과식초900ML - PET BOTTLE"|2|5500
"20131208"|"2030BC"|"8801007168746"|"010101"|"식초"|"씨제이제일제당 백설국내산사과로만든2배사과식초500ML - PET BOTTLE"|2|3100
```

Figure: 스캐너 데이터

스캐너 데이터의 결측치

- 상품의 판매 정보가 없는 경우 로그가 기록되지 않음
- 상품 개수 (285,000) X 상점 수(3,241) X 주 수 (352) $\simeq$ 3천 3백억
- 실제 로그 개수 8.4억은 가능한 기록의 **0.3%**에 불과함

- 결측치가 완전 무작위로 발생하지 않는 경우 관측 자료만을 이용한 분석이 실제 결과와 동떨어질 수 있다.
- 결측치를 효과적으로 보간하는 알고리즘이 필요 - **Factorization Machine**

자료 전처리

- 각 상품군 별로 나누어서 로그를 저장 (상품군: 188개)
- 유사한 품목에 대해 포장이 다르거나 용량이 다른 경우 상품명을 통일 (형태소 분석 알고리즘 적용)
  - CJ 사과식초 900ml / 2개 / 5,500원 + CJ 사과식초 500ml / 2개 / 3,100원 $\rightarrow$ CJ 사과식초 / 2800ml / 3.07 (원/ml)
  - 상품 수가 21만건으로 줄어듦

결측치 보정

- 각 주별로 각 상품의 판매 용량(Q)과 용량당 판매 금액(P)을 저장
  - $Q_{c,k,w,s}$: c번째 상품군의 k번째 상품이 w주에 s상점에서 팔린 총 판매 용량
  - $P_{c,k,w,s}$: c번째 상품군의 k번째 상품이 w주에 s상점에서의 용량당 판매 금액
- O: $(Q_{c,k,w,s}, P_{c,k,w,s})$가 모두 관측된 인덱스의 집합
- 결측치 보간: $\{Q_{c,k,w,s}, P_{c,k,w,s})|(c,k,w,s) \in O\}$를 이용해서 관측되지 않은 인덱스 $(c,k,w,s) \in O^c$에 대해 $(Q_{c,k,w,s}, P_{c,k,w,s})$값을 예측

행렬 분해 모형

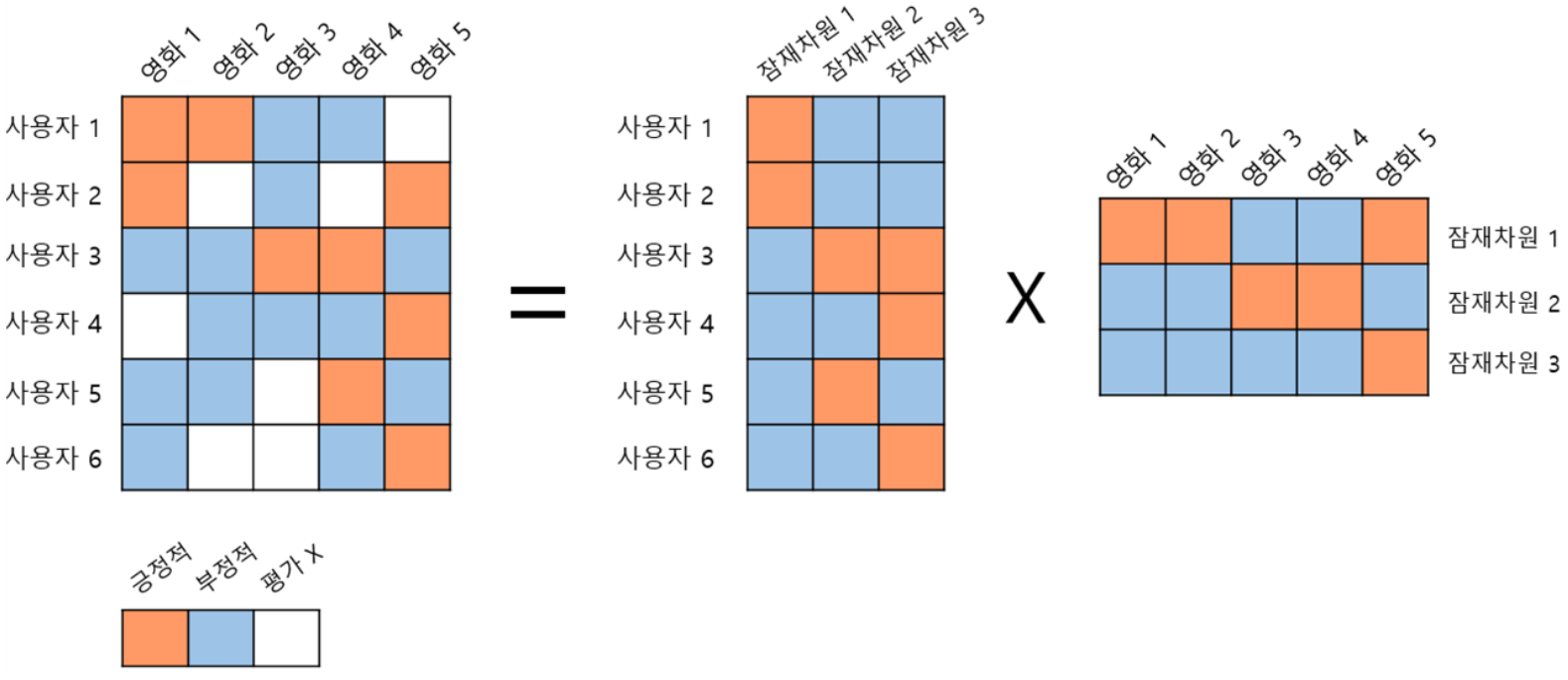
- $n \times m$ 크기의 자료 행렬 A가 저차원의 두 잠재 요인 행렬 $U \in \mathbb{R}^{n \times d}, V \in \mathbb{R}^{m \times d}$의 곱으로 표현 가능하다고 가정
- 전체 $n \times m$개의 자료를 $(n+m) \times d$개의 값으로 표현하므로 $d << m, n$이면 자료를 효율적으로 표현 가능
- U의 i행 벡터 u_i와 V의 j행 벡터 v_j에 대해 $A_{ij} = u_i^\top v_j$로 표현
  $\Rightarrow$ U의 각 행은 A의 행의 정보를 , V의 각 열은 A의 열의 정보를 담고 있다고 할 수 있다.

- U, V의 추정:

$$\sum_{A_{ij}:observed} \left(A_{ij} - u_i^\top v_j\right)^2 + L(U, V)$$

  를 최소로 하는 값으로 추정
  - $L(U, V)$: U, V에 대한 벌점 함수 (주로 L2 penalty 사용)

행렬 분해 모형


영화 1
영화 2
영화 3
영화 4
영화 5
사용자 1
사용자 2
사용자 3
사용자 4
사용자 5
사용자 6
=
잠재차원 1
잠재차원 2
잠재차원 3
X
긍정적
부정적
평가 X

행렬 분해 모형

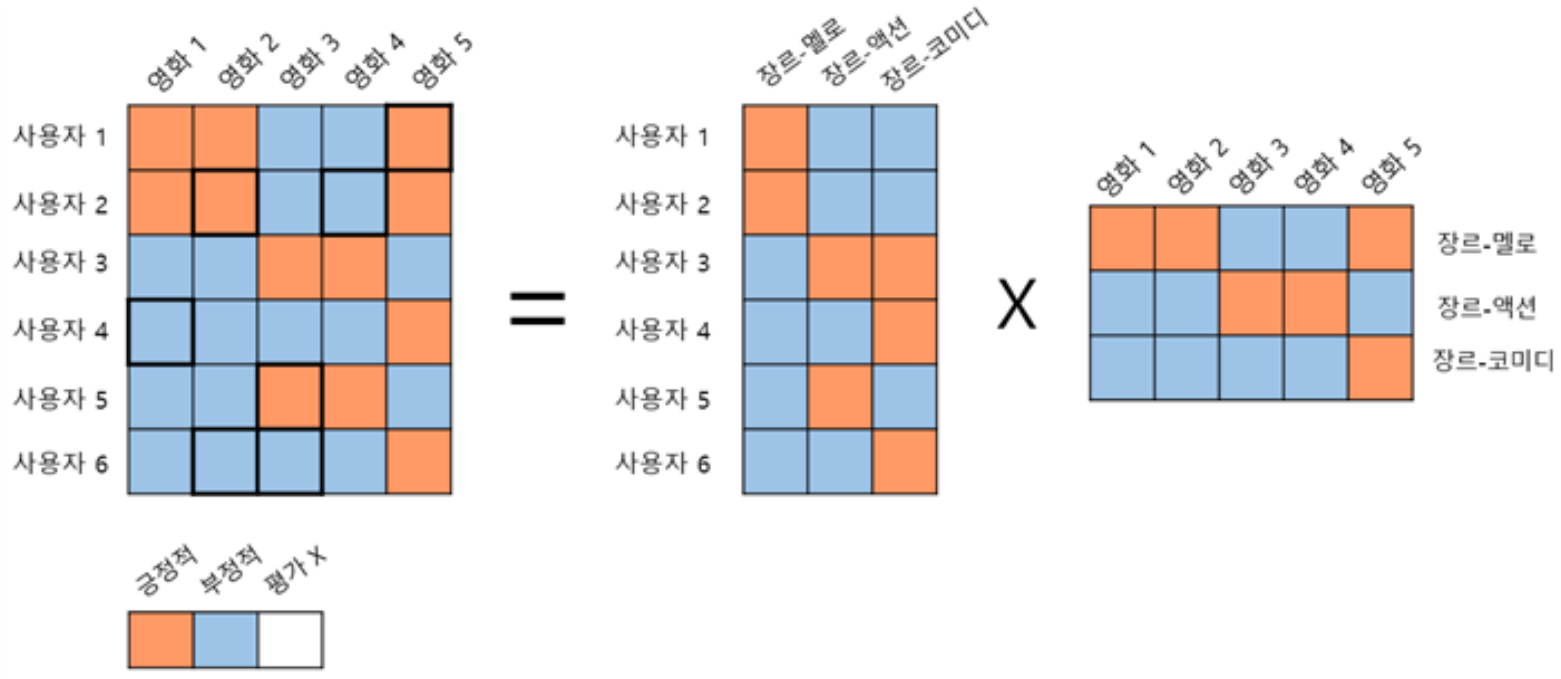


Figure: 행렬 분해 모형을 통한 결측치 보정. 검정 박스 부분이 행렬 분해 모형을 통해 채워진 결측치이다.

결측치 보간 모형으로써의 행렬 분해 모형

- 각 상품에 대해서 (상점,주)별 판매액 행렬과 판매량행렬에 행렬분해 방법 적용.
- 문제점: 상품이 너무 많다 (21만개).
- 해결책: Factorization machine알고리즘을 이용하여 (같은 상품군에 있는) 다수의 상품 에 대해서 동시에 분석

Factorization machine(FM)

- 회귀분석 문제: 독립변수 $x_1, \cdots, x_p$를 이용해 y를 예측
- 교호작용을 고려하는 모형

$$f(x_1, \cdots, x_p) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i<j} \gamma_{ij} x_i x_j$$

  모수가 p^2개 필요하기 때문에 분석에 어려움이 따른다.
- Factorization machine: 교호 작용 계수 γ_{ij}를 각 독립변수에 대한 잠재 변수 z를 이용해 표현

$$\gamma_{ij} = z_i^\top z_j, \quad z_1, \cdots, z_p \in \mathbb{R}^d$$

- 모형의 추정: MSE + 벌점함수를 최소로 하는 모수 추정

Factorization machine과 향렬분해

- 범주형 입력변수가 두개의 회귀모형을 고려
- 첫번째 변수는 행번호, 두번째 변수는 열번호
- 두개의 범주형변수를 가변수로 바꾼후 FM을 적용하면 행렬분해와 같음.
- FM는 행렬분해를 다차원 텐서 데이터로 확장할 수 있는 길을 제공함

스캐너 데이터와 FM

- 상품군별 스캐너 자료를 3차원 (상품명, 상점, 주) 텐서 데이터로 만들고 FM적용
- 같은 상품군에 대해 다음과 같은 잠재 벡터를 고려
  - $z_{c\cdot\cdot} \in \mathbb{R}^d, \beta_{c\cdot\cdot} \in \mathbb{R}$: c번째 상품의 잠재벡터, 일차항
  - $z_{\cdot w\cdot} \in \mathbb{R}^d, \beta_{\cdot w\cdot} \in \mathbb{R}w$번째 주의 잠재 벡터, 일차항
  - $z_{\cdot\cdot s} \in \mathbb{R}^d, \beta_{\cdot\cdot s} \in \mathbb{R}$: s번째 상점의 잠재 벡터, 일차항
- c번째 상품의 w번째 주의 s번째 상점에서의 예측값:

$$\beta_0 + \beta_{c\cdot\cdot} + \beta_{\cdot w\cdot} + \beta_{\cdot\cdot s} + z_{c\cdot\cdot}^\top z_{\cdot w\cdot} + z_{\cdot w\cdot}^\top z_{\cdot\cdot s} + z_{\cdot\cdot s}^\top z_{c\cdot\cdot}$$

방법론간 비교

- 식초 상품군과 스낵 상품군 (약 3천만 건의 로그를 가짐. 상품군 중 최다)에 대해 분석을 진행
- 비교 방법론: 선형 회귀, 행렬 분해, FM
- 관측치의 70%만을 실제로 관측했다고 가정하고, 나머지 30%를 얼마나 잘 예측했는지 MSE를 계산

상품군	판매 정보	σ^2	선형 회귀	행렬 분해	FM
식초	$P_{c,k,w,s}$	32.4	22.6	20.5	13.2
"	$Q_{c,k,w,s}$	7e+9	3.9e+9	5.5e+9	4.6e+9
스낵	$P_{c,k,w,s}$	1.69e+5	1.70e+5	1.69e+5	1.72e+5
"	$Q_{c,k,w,s}$	8.7e+7	5.7e+7	6.7e+7	7.5e+7

Table: 각 상품군에서 결측치 보간 방법론에 따른 예측값의 MSE. σ^2은 각 상품군에 속하는 검증자료의 분산이다.

결측치 보정에 따른 변동 변화

- 결측치 보간을 통해 완전한 텐서 자료 $P_{c,k,w,s}, Q_{c,k,w,s}$를 얻으면 이를 통해 물가 지수를 산출할 수 있다.
- 상품 (c, k)의 w주의 물가지수 $P_{c,k,w}$는 가중치가 $Q_{c,k,w,s}$인 $P_{c,k,w,s}$의 가중 평균으로 계산 가능하다.

$$P_{c,k,w} = \frac{\sum_s Q_{c,k,w,s} \cdot P_{c,k,w,s}}{\sum_s Q_{c,k,w,s}}$$

- 상품군 c의 물가지수$P_{c,w}$는 상품 (c, k)의 물가지수의 가중평균으로 고려할 수 있다.

$$P_{c,w} = \frac{\sum_s Q_{c,k,w} \cdot P_{c,k,w}}{\sum_s Q_{c,k,w}}$$

이 때 $Q_{c,k,w} = \sum_s Q_{c,k,w,s}$로 각 주에 상품 (c, k)가 팔린 총량을 의미한다.

결측치 보정에 따른 변동 변화

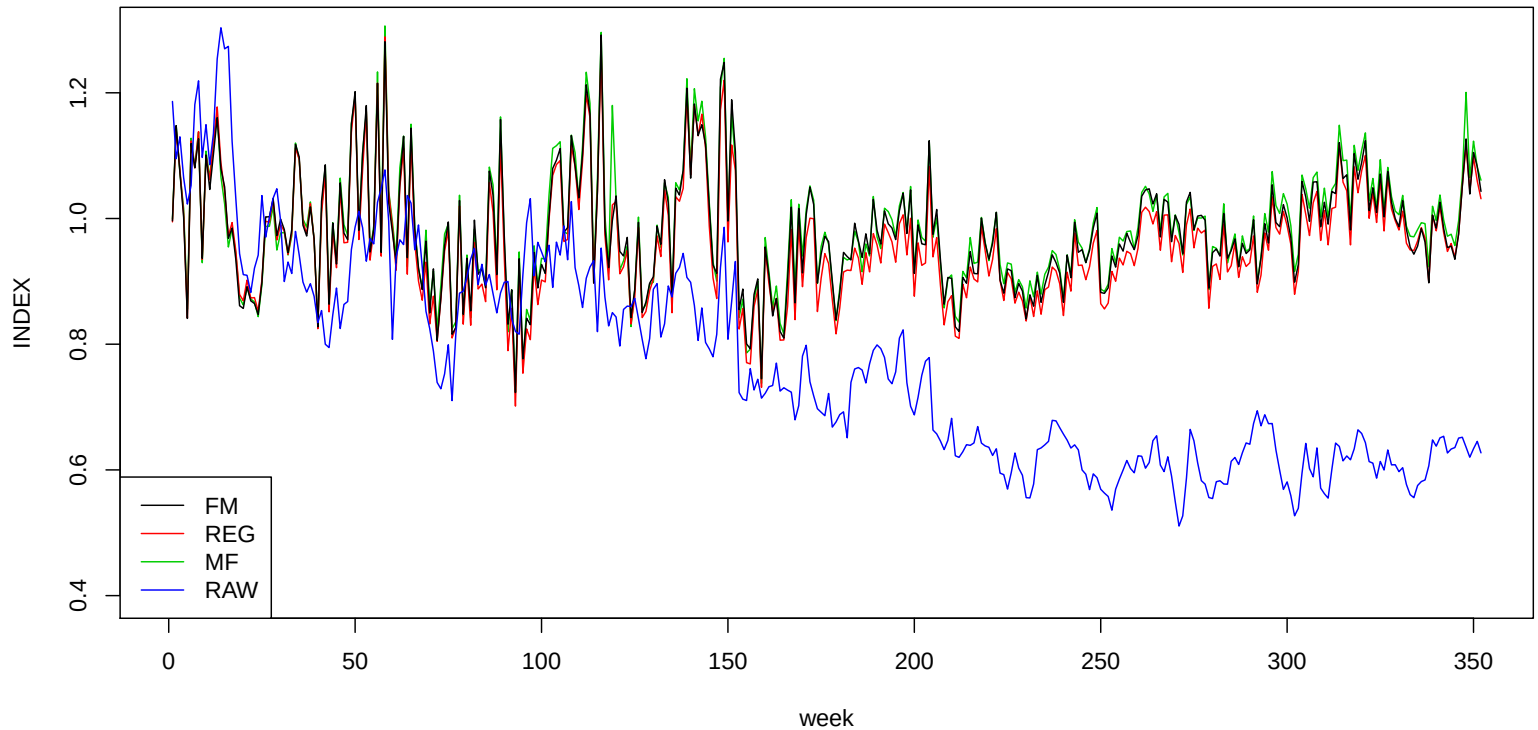


Figure: 식초 상품군에서의 물가 지수. 결측치 보정 방법론에 따라 다른 색으로 표현.

결측치 보정에 따른 변동 변화

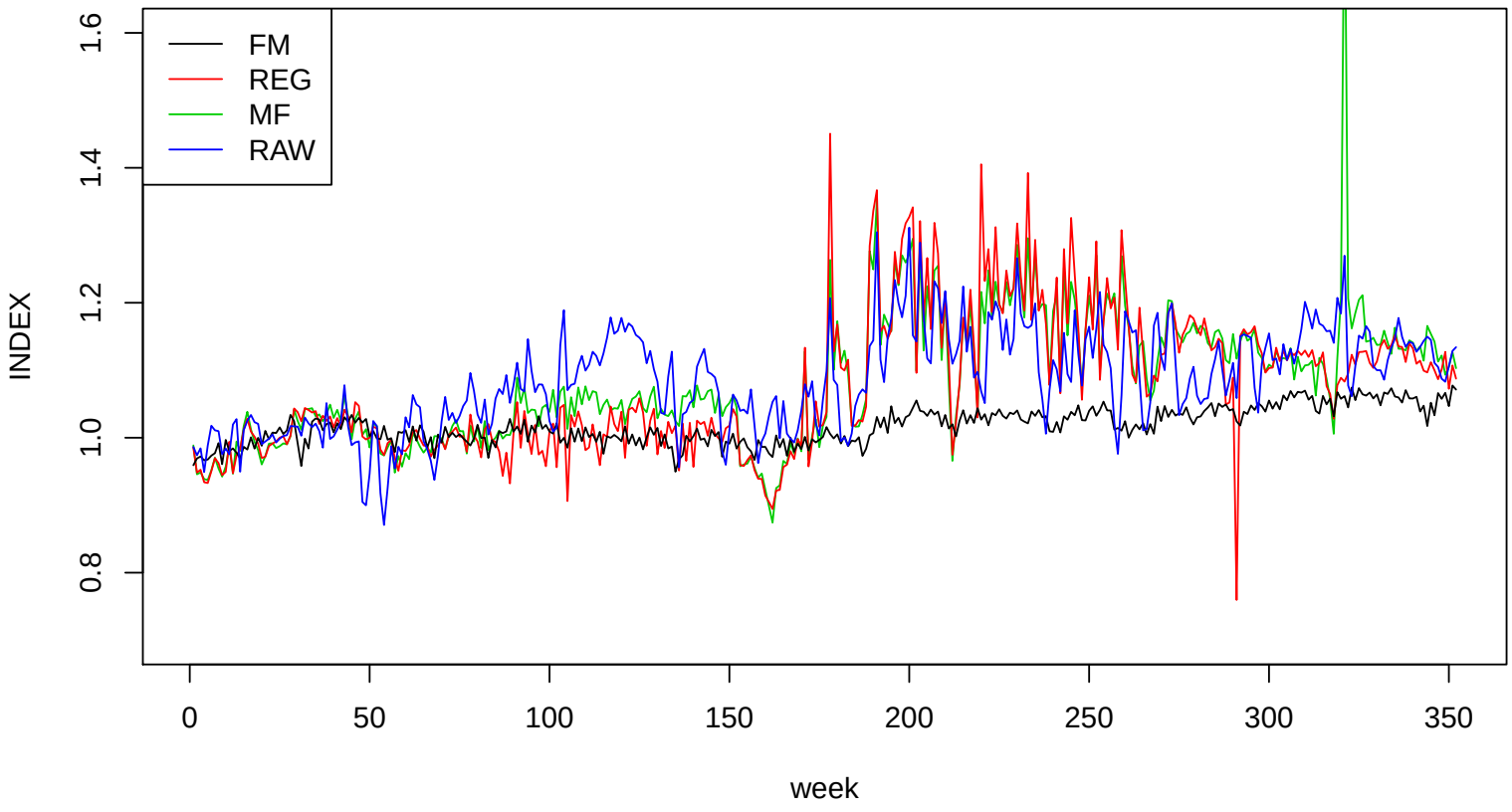


Figure: 스낵 상품군에서의 물가 지수. 결측치 보정 방법론에 따라 다른 색으로 표현.

분석 정리

- 행렬 분해 기법과 FM 모형을 통해 결측치를 보정 후 물가 추이 판별
- 물가지수의 국소적 변동을 줄임으로써 실제 자료만을 이용하는 것보다 결측치를 보정했을 때 거시적 물가 변동 파악에 도움이 됨

- 해결해야할 문제
  - 판매량이 0인 것과 판매량이 결측치인 것이 구분이 안되는 경우가 있슴
  - 신장개업하거나 폐업한 마트나 슈퍼가 다수 있는 것으로 파악되지만 정보가 부재
  - 스캐너 데이터 수집에 좀더 많은 노력이 필요함.

감사합니다.