다수준 회귀를 이용한 시간단위 군집화 모델:

물류산업의 파렛트 이동량 시계열 예측을 중심으로

Time Unit Clustering Model Using Multilevel Regression:

View from Time Series Analysis of Pallet Movement Amount of Logistics Industry

문현지(Hyunji Moon), 학부생, 서울대학교 공과대학 산업공학과, 010-5467-1667

e-mail: hjmoon0710@snu.ac.kr

서울시 관악구 관악로 1 서울대학교, Tel: 02-880-7172

국문초록

데이터 수집 및 처리 기술이 발전하면서 빅데이터 분석 및 활용 노력이 늘고 있다. 데이터는 더 높은 빈도와 다양성으로 수집되고 있으며, 분석은 더욱 세분화된 하위집단을 대상으로 시도된다. 시계열 자료에서 이러한 경향은 복잡한 계절성과 다양한 설명변수와 같은 새로운 데이터 특성을 시사한다. 시간축 혹은 시간축과 수직한 방향으로 예측이 세분화되는 것 또한 변화의 일부다. 그러나 많은 시계열 예측 모델은 구조적 한계로 인해 이런 새로운 데이터 속성을 반영하지 못하여 낮은 정확도의 예측을 한다. 본 논문은 시간 단위 군집화 모델을 제안한다. 이 모델은 시간 단위를 기준으로 시계열 데이터를 분류한 후 단위별로 분리된 시계열 집단들을 군집화한다. 또한 복잡한 계절성은 푸리에 방식으로, 다양한 설명변수는 베이즈 일반화된 선형 모델로, 세분화 집단의 특성은 계층적 모델로 모델링한다. 이는 기존 시계열 모델의 한계를 개선하고 더 높은 정확도로 예측값을 산출한다.

키워드: 물류 시계열 예측, 군집별 예측, 다수준 회귀분석

1. 서론

1.1 문제 정의

데이터 수집장치와 저장장치의 성능이 증가하고 가격이 낮아지면서 데이터를 수집하고 분석하는 비용이 크게 감소했다. 이에 물류산업의 데이터는 두 가지 측면에서 크게 변했다. 우선 데이터의 계절성이 복잡해졌다. 계절성의 주기가 길어졌으며 대부분의 시계열이 다중 계절성을 가지게 되었다. 이는 데이터 처리 비용의 감소에 따라 처리빈도가 증가했기 때문이다. 그러나 ARIMA (Autoregressive Integrated Moving Average), ETS (Exponential Smoothing)와 같은 기존 시계열 분석 방법들은 대체로 단기적인 패턴 분석에 장점을 가지므로 현재 상황을 분석하는데 한계를 보이고 있다.

또한 데이터의 구조가 세분화되고 있다. 과거에는 몇 개의 제품이 팔렸는지에 대한 데이터만을 수집했다면, 현재에는 몇 개의 제품이 어디에서 누구에게 팔렸는지까지 수집한다. 이는 예측모델에서 설명 변수로 이용가능한 데이터가 많아지고 예측문제가 복잡해짐을 의미하기 때문에 회귀 모델을 사용하는 것이 유리하다고 할 수 있다. 따라서 다수준 회귀모델을 통해서 데이터를 체계적으로 분석하면 좋은 예측결과를 얻을 수 있는 환경이 조성되었다.

물류 데이터에서는 업무일정 등에 영향을 받는 계절성이 관찰된다. 물류는 수요와 공급을 결정하는 사람들의 영향을 많이 받는데, 이들이 업무일정에 따라 의사결정을 내리기 때문이다. 이 의사결정자들의 행동은 연, 분기, 월, 주 등 달력에서 관찰되는 주기성을 보인다. 근무하는 평일이 휴일인 주말보다 물동량이 많다는 사실이 대표적인 예다. 그러나 기업과 산업마다 업무패턴이 다르므로 한 데이터에서 관찰한 계절성 패턴을 다른 데이터에 적용하기 어렵다. 해당 데이터로부터 계절성 패턴을 학습하는 등의 처리가 필요하다.

본 논문의 실험에 사용하는 데이터는 파렛트 대여 산업의 물류데이터이다. 실험대상 기업의 물류 구조는 일반적인 구조인 Figure 1과는 조금 다르지만 Figure 2 구조에서도 각 물류들은 서로에게 영향을 준다. 이 점에서 일반적인 구조와 크게 다르지 않다. 파렛트는 물류라는 추상적인 흐름을 명확하게 해주는 운송 매개체이다. 매일 수십만 종류의 다양한 제품들이 파렛트 위에 실려 이동된다. 많은 논문들에서 파렛트를 하나의 제품이 아닌 다른 제품의 운송을 위한 수단으로 본다. 하지만 본 논문에서는 파렛트를 다른 제품을 이동시키는 수단이 아닌 관리되어야 하는 하나의 제품의 관점으로 파렛트의 물류를 분석한다.

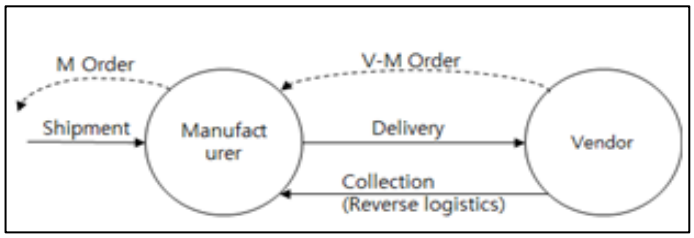


Figure 1. 일반적인 물류 흐름

파렛트 대여산업의 기업(Pallet Center)은 생산공장(Manufacturer)은 물론 판매처(Vendor)와도 빈번한 교류가 있어야 한다. 파렛트 대여 산업의 물류에는 각각 조달물류, 판매물류에 해당하는 파렛트 센터에서 생산공장으로, 생산공장에서 판매처로의 물류가 있다. 일반적인 산업의 물류 흐름과 차이가 나는 부분은 회수물류다. 일반 제품들의 회수물류는 판매처에서 생산공장으로 돌아오는 흐름이지만, 파렛트 대여산업에서 회수물류는 판매처에서 파렛트 센터로 돌아오는 흐름이다. 조달물류, 판매물류, 회수물류는 각각 Figure 2에서 Shipment, Delivery, Collection에 해당한다.

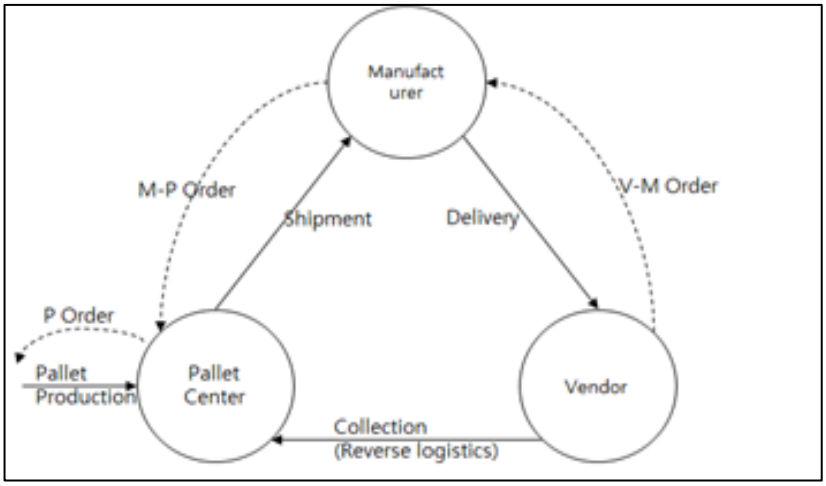


Figure 2. 실험대상 기업의 물류 흐름

파렛트 대여산업의 데이터를 사용하여 실험을 한 이유는 파렛트 대여산업이 물류산업을 대표할 수 있다고 판단했기 때문이다. 전자산업, 식품산업, 비료산업 등을 비롯한 수많은 산업의 기업들의 물류 흐름에 대한 데이터가 모두 파렛트 데이터에 포함된다. 따라서, 파렛트 데이터를 통해서 산업군별, 기업별 물류 흐름의 패턴을 파악할 수 있기 때문에 실험 데이터로서 적합하다.

1.2 연구 동기 및 공헌

시계열 데이터의 빈도 증가와 다양성 증가에 맞추어 시계열 분석 기술도 변해야 한다. 현재까지도 많이 사용되는 시계열 분석 방법론들에는 한계가 존재한다. 대표적인 시계열 분석 방법론인 ARIMA와 ETS는 1950년대 이전에도 사용되던 방법들이다. 이들이 안정적이고 신뢰할 수 있는 모델들임은 분명하지만, 모델들의 구조적 한계로 인해 길어진 계절성 주기와 다중 계절성을 모델에 반영하기 어렵다. 따라서 새로운 시계열 분석 방법론들이 개발되어야 하며 산업 현장에서는 이런 기술의 변화를 빠르게 수용해야 한다. 그럼에도 불구하고 현실에서는 기본적인 회귀모델도 이용하지 않고 실무자의 직관을 기반으로 주먹구구식 예측을 하고 있다. 이에 따른 경제적인 손실은 측정할 수 없을 정도로 크다.

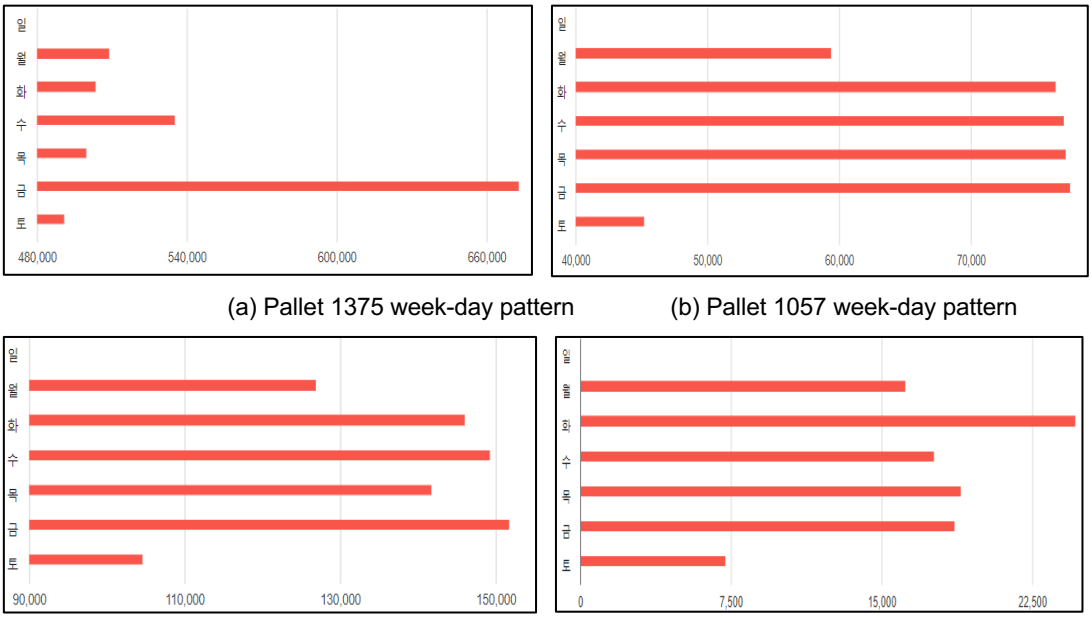
수요예측은 기업의 공급에 대한 의사결정의 근거가 되기 때문에 중요하다. 특히 시계열 데이터의 성분 중 계절성은 의사결정자에게 중요한 분석 결과이다(Terwiesch & Cachon, 2012). 물류산업의 많은 기업들이 다루는 공급망관리(Supply Chain Management) 분야를 예로 들어보자. 공급망관리의 목적은 불확실한 수요에 대응하여 이익을 최대화하고 비용을 최소화하는 공급계획을 세우는 것이다. 여기서 수요의 불확실성을 얼마나 줄이는가에 따라서 공급망관리 문제의 복잡도가 결정되기 때문에 수요예측이 매우 중요하다.

시계열 데이터의 새로운 특성과 수요예측의 중요성이 시간단위 군집화 모델(Time Unit Clustering, TUC)의 연구배경이다. 연구 중 본 논문에서 제안하는 모델을 국내의 한 파렛트 대여기업의 파렛트 발주, 입고, 회수량 예측에 적용했다. 해당 기업은 실무 담당자들의 경험에 따라 수요를 예측하고 있었으며, 평균 오차율이 약 30%이다. 이는 파렛트의 생산, 운송, 회수계획 수립에 크게 도움이 되지 않을 정도의 예측정확도이다. 본 연구에서 사용한 시간단위 군집화 모델을 통해 평균 오차율이 약 7%로 줄었다. 본 논문의 제시한 시간단위 군집화 모델을 통해서 물류산업의 다양한 기업들의 수요예측의 정확도를 높이고 해석가능한 결과를 제시함으로써 물류 운영의 효율성을 크게 개선할 수 있기를 기대한다.

2. 산업 및 기술 분석

2.1 물류산업과 파렛트

실험 데이터는 파렛트 대여산업의 한 물류업체(이하 A사)가 2010년 01월 01일부터 2018년 05월 31일까지 약 8년 동안 언제 어디에서 어디로 어떤 유형의 파렛트가 몇 개 이동되었는지를 일 단위로 수집한 자료이다. 고객의 정보보호를 위해 파렛트와 생산공자 및 판매처 등에 대한 정보는 코드로 암호화하여 표기한다. Figure 3은 네 가지의 유형에 따른 파렛트 이동량의 요일별 평균을 나타낸 것이다. 한 파렛트를 기준으로 보면 요일간 이동량의 차이가 존재함을 알 수 있다. 많은 고객이 휴일인 일요일과 일요일이 아닌 요일의 평균이 대략 1:200 정도로 차이가 나며, 각 요일 간에도 분명한 차이가 존재한다. 또한, 이런 요일별 차이의 패턴이 파렛트의 유형마다 다름도 알 수 있다. A사의 고객은 식품기업, 제관기업, 비료기업 등 다양한데, 분석결과 파렛트가 사용되는 산업과 주요 고객의 주문 방식 등에 따라 요일별 파렛트 이동량의 패턴이 달라졌다.



(a) Pallet 1375 week-day pattern (b) Pallet 1057 week-day pattern

(c) Pallet 1091 week-day pattern (d) Pallet 1817 week-day pattern

Figure 3. 네 가지 유형의 파렛트 이동량의 요일별 패턴

2.2 시계열 분석 기술

ARIMA와 같이 전통적인 시계열 분석 기법으로 추세성과 계절성이 동시에 나타나는 시계열 데이터를 수식 (1)과 같이 계절차분, 일차 차분하여 정상시계열로 변환시킨다. 그러나 ARIMA 모델은 분기 혹은 월 단위처럼 계절성 주기(seasonal period, 한 주기성을 관찰하기 위해 필요한 데이터 수)가 짧은 경우에만 적용 가능하다는 단점이 있다. 이 단점은 최대 계절성 주기가 24로 제한된 ETS 모델에서도 발견된다(Hyndman & Athanasopoulos, 2018).

$$X_t = T_t + S_t + Z_t \quad (1)$$

물류시스템의 디지털화와 데이터의 수집 구조 발달에 따라 시계열 데이터가 짧은 시간 단위로 저장되고, 이에 기존 ARIMA 모델과 ETS 모델 등의 전통적인 시계열 방법론으로는 높은 예측정확도를 기대하기 어렵다. 예를 들어 30분 단위로 수집되는 데이터는 48의 주기를 가지지만, ARIMA 모델과 ETS 모델은 긴 계절성 주기의 표현이 어렵다는 구조적 한계로 인해 1일 주기성이 누락된 예측값을 제공한다.

이에 Taylor & Letham (2017)은 유연성(flexibility), 결측치가 있는 시계열의 수용가능성, 빠른 적합(fast fitting), 해석 가능한 모수값 등의 장점을 가진 프로펫 예측모델(prophet forecasting model)을 제안했다. 이 모델은 Harvey & Shephard (1993)가 제안한 구조화된 시계열, 즉 분해 가능한 시계열 모델(decomposable time series model)을 기반으로 시계열 y(t)를 수식 (2)와 같이 표현한다.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (2)$$

시계열 회귀는 시계열 y가 다른 시계열 x와 선형적 관련성이 있다는 전제 하에 y를 예측하기 위해 다른 시계열 x를 이용하는 방식이다(Hyndman & Athanasopoulos, 2018). 이때 y를 예측 변수(forecast variable), x를 설명변수(explanatory variable)라 한다. 시계열 회귀를 이용하면 앞서 1.1에서 설명한 물류 구조도에서 각 물류흐름의 파렛트 이동량을 서로의 예측변수로 이용할 수 있으며 이를 통해 예측정확도를 높일 수 있다.

2.3 군집별 예측 기술

군집별 예측은 일차적으로 모집단을 군집화 후 각 군집마다 독립적인 예측모델을 만드는 방식이다. 이를 군집 후 예측(cluster-then-predict) 혹은 군집별 예측(cluster-wise prediction)이라 한다. 이 방식을 이용하여 Bertsimas (2015)는 환자의 심장마비 분류예측정확도를 개선하였다. 그는 심장마비에 이르는 환자들의 건강패턴이 다양하므로, 일차적으로 환자들의 데이터를 군집화하여 유사한 환자들은 모은 후, 각 군집을 기반으로 모델을 만드는 것이 더 정확하다고 설명했다.

Venkatesh (2013)은 한 지역의 현금자동인출금기를 시계열 패턴을 이용하여 군집화한 후 군집별로 독립적인 예측을 했다. Venkatesh가 사용한 군집화 기준은 한 현금자동인출금기의 요일별 현금인출량으로, 요일효과(day-of-the-week effect)를 이산화시킨 길이 7인 수열과 SAM(sequence-alighment method)을 이용해 군집화를 진행했다. 이를 통해 군집화 선처리 과정 없이 전체 ATM으로 예측모델을 만들었던 기존방식에 비해 더 정확한 예측을 했으며, 군집화 결과를 바탕으로 운영자들은 현금 보충 계획을 더 효율적으로 짤 수 있다고 설명했다.

2.4 다수준 회귀모델 기술

McElreath (2016)는 다수준 구조로 모델링을 하지 않을 이유가 없는 이상 다수준 구조로 모델링하는 것이 더 좋고 정확하다고 주장한다. 2.2의 군집별 예측 기술은 모든 군집이 독립적인 모수를 학습한다. 이에 비해, 다수준 회귀 기술은 군집 간 공통모수 학습(complete-pooling)에서부터 독립적인 모수(no-pooling) 학습까지의 넓은 스펙트럼을 제공한다. 현재 베이지안 추론(Bayesian Inference) 분야에서 활발히 연구되는 계층모델도 이 스펙트럼에

포함된다. 계층모델에서는 군집들이 모수의 분포를 결정하는 공통초모수를 학습(partial-pooling)한다. 따라서 2.2의 군집별 예측 기술은 2.3의 다수준 회귀 기술의 구체적 형태라 볼 수 있다.

3. 시간단위 군집화 모델

3.1 필요성

큐빅 스플라인(cubic spline)과 같은 비선형 추세 모델들은 데이터 적합은 잘 되지만, 실제 예측정확도가 떨어지므로 예측구간에서는 선형 추세 모델이 더 정확하다(Hyndman & Athanasopoulos, 2018). 선형 추세 모델 중에는 구간별 선형 추세(piecewise linear trend) 방식이 효과적이다 (Taylor & Letham, 2017).

그러나 추세는 시간 단위의 영향을 받는다. 예를 들어 경기활성화로 소비가 증가해도 회사 밀집지역인 마포구 식당의 저녁고객의 증가추세는 요일별로 차이가 있을 수 있다. 월~금요일 저녁의 기울기 변화가 토, 일요일의 기울기 변화보다 더 클 것으로 예상된다. 금요일의 밤 문화로 금요일의 증가추세가 가장 클 수도 있다. 그러나 기존 구간별 선형 추세 방식에서는 이런 구분 없이 이 경우 '일' 단위에 해당하는 7개의 그룹들이 같은 추세를 가지는 방식으로 모델링했다.

마찬가지로 계절성 성분은 푸리에 계열을 주로 이용하는데, 시간 단위별로 푸리에 계수가 다를수 있음에도 모든 시계열 성분들이 동일한 푸리에 계수를 가지도록 모델링된다. 이는 앞서 설명한 다수준 모델에서 모델을 '군집 간 공통모수 학습(complete-pooling)'으로만 국한시킨다. 세부 군집들이 모여 하나의 큰 군집을 형성하는 구조에서 세부 군집간의 차이가 클 때 '군집 간 공통모수 학습'방식은 좋지 않다(Gelman, 2013). 이를 개선하기 위해 본 논문에서 제안하는 방식은 시계열을 시간단위로 이산화한 후 유사한 요약통계량을 가지는 시간단위로 군집화 한다. 그 후 군집별로 독립적인 모수를 학습한다. 자세한 설명은 3.2에서 진행한다.

3.2 세부 단계

본 논문에서는 과거 요일별 시계열의 요약통계량을 바탕으로 군집화한 후 군집마다 다른 모수를 갖는 모델을 설정하는 방식을 적용해보고자 한다. 전체 데이터를 바탕으로 예측모델을 적합시키지 않고, 시계열을 시간단위별로 이산화를 하여 군집화를 한 후 각 군집별로 독립적인 예측모델을 만든다. 이 방식을 '시간 단위 군집화(Time Unit Clustering)'라 정의한다.

시간단위 군집화 모델은 기본적인 토대를 Tayor & Letham (2017)의 프로펫 예측모델에 두고 있다. 프로펫 예측모델은 유연성, 결측치 수용가능성, 빠른 적합성, 모수의 해석가능성 등의 장점을 가진 시계열 분석 기법이다. 이 모델은 Harvey & Peters가 제안한 분해 가능한 시계열 모델을 따라 시계열 데이터를 추세, 계절성 등의 성분으로 분해한다. 해당 추세성분은 구간별 선형모델을, 계절성 성분은 푸리에 모델을 이용하며, 베이지안 방식을 이용하여 데이터로부터 추세와 계절성 성분을 결정하는 모수를 학습한다.

시간단위 군집화 모델은 기존 프로펫 예측모델에 '군집화 후 예측' 방식을 적용한다. 프로펫 예측모델에서 모든 시계열 자료를 바탕으로 추세와 계절성 성분의 모수를 학습하였다면, 시간단위 군집화 모델은 시계열을 시간단위로 이산화하여 군집화한 후 각 군집들이 독립적으로 모수를 학습한다. 이런 군집화 선처리 과정이 필요한 이유는 시간 단위들 간의 성질이 다르기 때문이다. 이때 시간 단위란 '일', '주', '월' 과 같이 시계열 데이터 생성에 영향을 주는 이산적인 시간 단위들을 말한다. 실험에 이용된 물류 파렛트 이동량 시계열 데이터를 예로 들면, 평일의 이동량과 일요일의 이동량은 그 크기가 200배 이상 차이 난다. 이에 본 논문의 실험에서는 시간 단위 '요일'에 대해, 시계열을 요일을 기준으로 7개로 이산화했다. 그 후 k-means 군집화 방식을 적용하였다. 7개의 요일을 각각 1개에서 7개 군집개수(k)에 따라 군집화하여 7개의 군집화 집합을 생성했다. 그 후 교차검증을 통해 가장 교차검증 오차가 적은 군집의 개수(k)를 선정하여 최종 예측모델로 삼았다. 전 과정을 Figure 4에 표현했다.

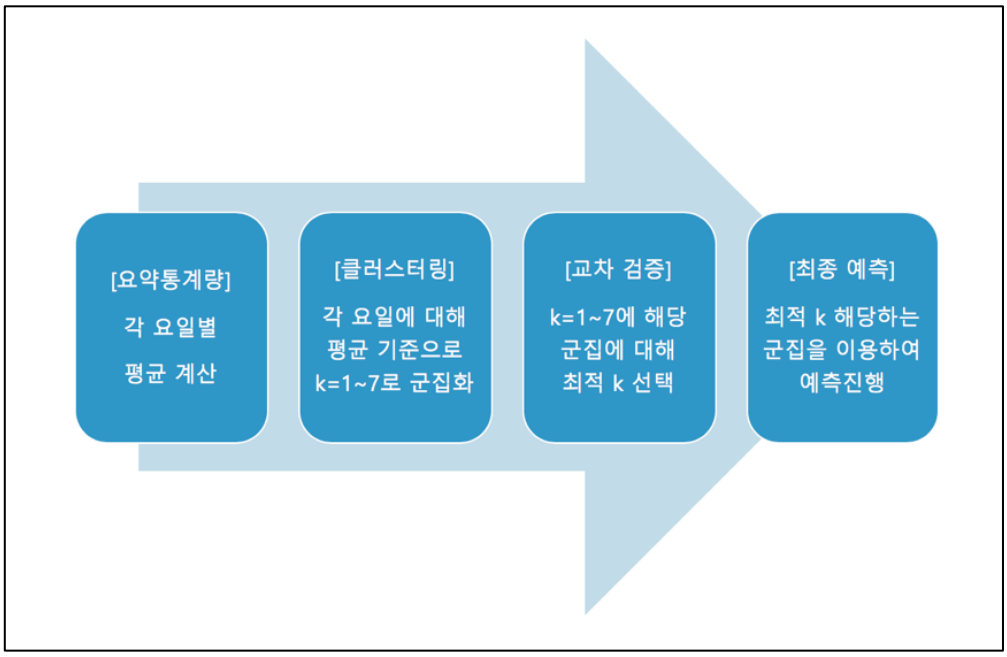


Figure 4. 시간단위 군집화 모델의 진행단계

4. 실험

4.1 기본 비교 모델

Taylor & Letham(2017)은 프로펫 모델이 ARIMA, 지수평활법, SNAIVE 모델, TBATS 모델보다 예측력이 뛰어남을 실험을 통해 보인바 있다. 따라서 프로펫 모델과 본 논문의 방식만 비교해도 충분하다. 프로펫 모델의 수리적인 시계열 구조는 TUC 모델과 같다. 차이점은 군집 별 예측이다. 프로펫 모델에서는 모든 요일이 동일한 추세와 계절성 모수를 공유하는 반면에 TUC 모델은 동일 군집에 속하는 요일들만 추세와 계절성 모수를 공유한다.

4.2 예측 정확도 측정 기준

예측 정확도의 측정 기준으로 보편적으로 사용되는 것으로는 MAP(Mean Absolute Error), MAPE(Mean Absolute Percentage Error), MSE(Mean Squared Error), MSPE(Mean Squared Prediction Error)이 존재한다. MAP는 실제값의 크기와는 상관없이 오차값의 크기가 중요한 경우에 사용하는 지표이고, MAPE는 오차값의 크기뿐만 아니라 실제값과의 상대적인 비율이 중요한 경우에 사용하는 지표이다. MSE는 오차값의 크기가 커질수록 비용이 증가하는 경우에 사용하는 지표이며 training data를 기준으로 측정하며, MSPE는 MSE와 마찬가지로 오차값의 크기가 커질수록 비용이 증가하는 경우에 사용하지만 test data를 기준으로 측정한다.

물류산업의 경우 오차값의 크기가 커질수록 비용이 증가하는 경우가 아니라, 오차값의 크기가 중요한 경우가 많기 때문에 위의 네 가지 지표 중 MAP가 가장 적합하다고 판단했다. 따라서 MAP를 기준으로 모델들의 예측정확도 우위를 판별하고자 한다.

4.3 실험 데이터

실험 데이터는 파렛트 대여업체 A의 파렛트가 2012년 01일 01일부터 2017년 12월 31일까지 언제 어디에서 어디로 어떤 유형의 파렛트가 몇 개 이동되었는지를 일 단위로 수집한 자료이다. 1815, 1041, 1627 코드로 표현되는 세 가지 유형의 파렛트를 실험 데이터로 사용한다. 이들은 각각 회사 거래량의 약 60%, 6%, 2%를 차지하는 유형들이기에 이동량 규모에 따른 모델의 적합도도 비교할 수 있게 설계되었다.

3.2의 세부 단계를 따라 실험이 진행되었다. 학습 데이터(train dataset)은 2012년 01일 01일부터 2017년 12월 31일까지의 데이터를 이용했으며 교차검증 데이터(cross-validation dataset)는 2017년 01월 01일에서 12월

31일까지의 12개월을 12등분으로 나누어 이용했다. 일반적인 시계열 교차검증방법을 따랐다(Arlot & Celisse, 2010). 교차검증을 통해 최적의 군집개수(k)를 선택한 후 시험 데이터(test dataset)인 2018년 01월 01일부터 2018년 01월 31일까지를 예측했다. 프로펫 모델과 시간단위 군집화 모델의 비교는 시험 데이터의 예측오차 비교를 통해 이루어졌다.

5. 실험 결과

Table 1 에서는 파렛트 3 개의 요일별 평균 이동량을 구했다.

Table 1. 파렛트 1815, 1041, 1627의 요일별 평균 이동량

파렛트 종류	월	화	수	목	금	토	일
1815	35,107	33,339	33,516	33,533	33,839	25,831	951
1041	5,252	5,590	5,946	5,938	6,912	5,501	6
1627	1,070	1,499	1,239	1,386	1,474	954	1

Table 1의 요약통계량을 바탕으로 군집 개수 1에서 7까지의 k-means 군집화를 진행하였고, 그 결과는 Table 2에서 확인할 수 있다. 월요일에서 일요일을 각각 0에서 6으로 표시했다.

Table 2. 파렛트 1815, 1041, 1627의 군집화 결과

군집화 개수	Pallet 1815	Pallet 1041	Pallet 1627
1	[0 0 0 0 0 0 0] ‘0123456’	[0 0 0 0 0 0 0] ‘0123456’	[0 0 0 0 0 0 0] ‘0123456’
2	[1 1 1 1 1 0 0] ‘01234’, ‘56’	[0 0 0 0 0 0 1] ‘012345’, ‘6’	[0 0 0 0 0 0 1] ‘012345’, ‘6’
3	[1 1 1 1 1 0 2] ‘01234’, ‘5’, ‘6’	[0 0 0 0 2 0 1] ‘01235’, ‘4’, ‘6’	[2 0 2 0 0 2 1] ‘025’, ‘134’, ‘6’
4	[3 1 1 1 1 0 2] ‘0’, ‘1234, ‘5’, ‘6’	[0 0 3 3 2 0 1] ‘015’,’23’, ‘4’, ‘6’	[0 2 3 2 2 0 1] '05','134','2','6'
5	[3 1 1 4 1 0 2] ‘0’, ‘124’,’3’, ‘5’,’6’	[4 0 3 3 2 0 1] ‘0’,’15’,’23’, ‘4’, ‘6’	[4 2 3 2 2 0 1] ‘0’,’2’,‘134’,’5’,‘6’
6	[3 1 5 4 5 0 2] ‘0’,’1’,’24’,’3’,’5’,’6’	[4 5 3 3 2 0 1] ‘0’,’1’,’23’, ‘4’,’5’, ‘6’	[4 2 3 5 2 0 1] ‘0’,’14’,’2’,‘3’,’5’,‘6’

7	[3 1 5 4 6 0 2] ‘0’,’1’,’2’,’3’,’4’,’5’,’6’	[4 5 3 6 2 0 1] ‘0’,’1’,’2’,’3’,’4’,’5’,’6’	[4 6 3 5 2 0 1] ‘0’,’1’,’2’,’3’,’4’,’5’,’6’

Table 3에서는 각 군집 결과를 바탕으로 교차검증을 한 결과값을 명시하였다. 1815 파렛트는 군집 개수가 4, 1041은 군집 개수가 2, 1627은 군집 개수가 4일때 가장 적은 교차검증오차를 보였다.

Table 3. 군집화 개수에 따른 교차검증오차(MAE)

군집화 개수	Pallet 1815	Pallet 1041	Pallet 1627
1	7,123	1,276	457
2	5,214	**1,153**	421
3	4,453	1,172	416
4	**4,441**	1,176	**412**
5	7,949	1,163	413
6	4,577	1,160	414
7	4,575	1,173	416

Table 4에서는 교차검증 결과 최종 선택된 군집 개수에 해당하는 군집화 집합을 이용한 예측모델과 기존 프로펫 모델의 예측오차를 비교한다.

Table 4. 최종 선택된 군집화 집합과 기존 프로펫 모델의 예측오차(MAE) 비교

파렛트 종류	최종선택된 군집	프로펫 모델	TUC 모델
1815	‘0', '1234', '5', '6'	5,281	3,597
1041	'012345', '6'	1,056	1,054
1627	'05','134','2','6'	503	453

6. 결과 분석 및 결론

Table 1, 2에서는 각 파렛트마다 요일별 패턴이 다르며, 따라서 군집화를 했을 때 모두 다른 형태로 군집화가 진행됨을 알 수 있다. 패턴이 파렛트 유형별로 다르다는 것은 일반적인 군집화를 모든 시계열에 적용하기 어려우며, 본 논문에서 진행했듯이 데이터로부터 군집 패턴을 학습해야함을 알려준다. Table 4에서 알 수 있듯이 TUC 모델은 기존 프로펫 모델보다 약 10%~20% 예측오차의 개선이 있다. 이 개선은 실제 현업의 운영을 크게 효율화한다.

추가 연구 제안은 다음과 같다. 본 논문에서는 다루지 못했지만, 물류 구조도 특성 상 물류 흐름들이 서로의 설명변수가 되는 모델이 예측정확도 개선에 큰 효과가 있으리라 생각한다. 또한 본 논문에서 관찰한 파렛트 이동량의 요일별 패턴과 파렛트 유형 혹은 산업 간의 관련성 또한 좋은 연구 주제다. 특정 산업만의 전형적인 패턴을 찾을 수 있다면 더 정확한 예측이 가능하다. 마지막으로 다수준 회귀모델과 관련하여 다양한 연구가 가능하다. 본 논문에서는 다수준 모델의 스펙트럼 중 일부분만을 시도해보았으며 더 다양한 모델을 적용해볼 수 있다. 본문 2.4에서 언급한 공통초모수를 학습하는 모델이 좋은 예이다.

데이터를 저장하고 처리하는 기술의 발달로 인하여 물류산업에서 다루는 시계열 데이터의 빈도가 높아지고 종류가 다양해지게 되었다. 이에 따라서 시계열 데이터의 계절성의 주기가 길어지고 다중계절성이 나타난다. 전통적 ARIMA (Autoregressive Integrated Moving Average) 모델과 ETS (Exponential Smoothing) 모델의 장기적인 계절성과 다중 계절성을 분석하기 힘들다는 한계를 지니고, 기존 프로펫 모델은 공통모수 학습이라는 한계를 지녔다. 따라서, 본 논문에서는 다수준 회귀의 관점에서 시간단위 군집화 모델을 제안하여 이러한 한계점을 개선하였다.

7. 참고문헌

Arlot, S., & Celisse, A. (2010), A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., & Wang, G. (2008), Algorithmic prediction of health-care costs. Operations Research, 56(6), 1382-1392.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990), STL: A Seasonal-Trend Decomposition. Journal of Official Statistics, 6(1), 3-73.

De Livera, A., Hyndman, R., & Snyder, R. (2011), Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing, Journal of the American Statistical Association, 106(496), 1513-1527.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013), Bayesian data analysis. Chapman and Hall/CRC.

Hyndman, R. J., & Athanasopoulos, G. (2018), Forecasting: principles and practice. OTexts.

Hyndman, R. J., Khandakar, Y. et al. (2007), Automatic time series for forecasting: the forecast package for R, number 6/07, Monash University, Department of Econometrics and Business Statistics.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. & Grose, S. (2002), 'A state space framework for automatic forecasting using exponential smoothing methods', International Journal of Forecasting 18(3), 439–454.

McElreath, R. (2016), Statistical Rethinking: A Bayesian course with examples in R and Stan / Richard McElreath, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.(Texts in statistical science ; 122), Boca Raton: CRC Press/Taylor & Francis Group.

Harvey, A. C. & Shephard, N. (1993), Structural time series models, in G. Maddala, C. Rao & H. Vinod, eds, 'Handbook of Statistics', Vol. 11, Elsevier, chapter 10, pp. 261–302.

Terwiesch, C., & Cachon, G. (2012), Matching supply with demand: An introduction to operations management. Mcgraw-hill Education-Europe.

Venkatesh, K. V., Ravi, V., Prinzie, A., & Poel, D. (2013), Cash demand forecasting in ATMs by clustering and neural networks. European Journal of Operational Research, 2013.