석사학위논문

# 시계열 군집분석 기반
# 서울시 공공자전거 수요예측

# Forecasting public bicycle demand of seoul based on time series clustering

김 민 혁

한양대학교 대학원

2018년 2월

석사학위논문

# 시계열 군집분석 기반
# 서울시 공공자전거 수요예측

# Forecasting public bicycle demand of seoul based on time series clustering

지도교수 차 경 준

이 논문을 이학 석사학위논문으로 제출합니다.

２０１８년　2월

한양대학교 대학원

응용통계학과

김 민 혁

이 논문을 김민혁의 석사학위 논문으로 인준함

2018년 2월

심 사 위 원 장 : 최 정 순 (인)

심 사 위 원 : 박 영 선 (인)

심 사 위 원 : 차 경 준 (인)

한양대학교 대학원

# 목 차

# 표 목차

# 그림 목차

# 국문요지

## 시계열 군집분석 기반 서울시 공공자전거 수요예측

한양대학교 대학원
응용통계학과
김민혁

본 연구에서는 서울시에서 서비스를 제공하고 있는 공공자전거의 원시 대여 이력 자료에 대해 시계열 군집분석(time series clustering)을 진행한 후 기계학습 방법 중 하나인 서포트 벡터 회귀(support vector regression)를 이용해 수요를 예측하였다. 서울시 전체 공공자전거 수요를 예측하기 위해서 모든 대여소의 대여량의 예측값을 사용하는 모형과 시계열 군집분석을 이용해 수요 패턴이 같은 대여소들을 군집으로 묶어 각각의 군집마다 최적의 모형을 만들어 예측값의 합계를 비교하는 방법을 제안하였다. 시계열 군집분석을 사용한 모형과 사용하지 않은 기존 모형을 서포트 벡터 회귀를 이용해 비교한 결과 시계열 군집분석을 기반으로 한 수요 예측 서포트 벡터 회귀모형의 평균 제곱근 오차(root mean square error)와 평균 절대오차(mean absolute error)의 값이 12%정도 더 낮아졌음을 확인하였다. 시계열 군집분석에서는 사용하는 거리 계산방법별 결과를 비교해 최적의 군집모형을 선택하였고, 서포트 벡터 회귀모형은 커널별로 모수를 다르게 한 결과를 비교해 최적의 모형을 선택하였다. 이 결과를 통해 시계열 군집 분석을 통해 대여소를 군집으로 나누고, 예측 모형을 만들면 기존보다 대여량을 더 정확하게 예측할 수 있으리라 기대된다.

# 1장. 서론

## 1. 연구의 배경 및 목적

최근 이동수단의 발달 추세는 대기오염과 온실가스를 줄이는 친환경 이동수단으로 발달되고 있다. 특히 자전거는 환경을 저해하는 요인을 발생시키지 않고, 도시에서 짧은 거리를 빠르게 이동할 때 교통 체증을 피할 수 있는 장점을 가지고 있어 세계 주요 도시들은 공공 자전거 시스템을 긍정적으로 도입하고 있다. 또한 동시에 시민들의 건강 증진을 도모할 수 있음은 물론 도시 관광자원으로까지 확대될 수 있기 때문에 공공자전거의 수요는 점점 폭발적으로 증가하고 있다. 이에 대해 신희철 등 (2012)은 공공 자전거의 도입으로 인한 효과와 발전 가능성에 대해 공간적, 시간적, 내용적으로 나누어 필요성을 강조하였다. 국내에서는 2008년 창원시에서 처음으로 공공 자전거 시스템인 “누비자”를 선보였고, 대전시는 공공 자전거 시스템 “타슈”를 성공시켰다. 서울시도 2015년부터 공공 자전거 시스템 “따릉이”를 구축하였는데, 김동준 (2017)은 설문조사 결과로 서울시 공공 자전거 시스템의 공유 인지도와 만족도의 우수함에 대해 보도하여 서울시 공공 자전거의 수요는 점점 늘어날 것으로 전망하였다. 또한 과거와는 달리 4차 산업혁명 시대를 맞이하여 공공 자전거 시스템은 최첨단 시설들로 관리되면서 이용자들의 이용 이력이 없어지지 않고, 막대한 양의 자료로 축적하여 시간에 따른 이용자들의 수요 패턴을 분석할 수 있는 기반을 마련하였다.

그러나 공공 자전거의 필요성에 비해 국내에서 수요 예측에 관한 연구는 상대적으로 드물다. 노승윤 등(2014)은 대전시의 공공 자전거 시스템의 이용패

턴 및 수요예측을 다중회귀분석(multiple regression analysis)을 사용하여 분석하였다. 연령층의 비율, 버스승하차, 수변공간까지의 거리, 공원여부 등을 변수로 사용해 이용객수를 예측하였는데, 주변 환경 변수에 초점을 맞춘 연구를 시행하였다.

민지원 등(2017)은 기계학습 방법(machine learning)인 랜덤 포레스트(random forest)를 이용해 대전시 공공 자전거 수요 예측을 하였으나, 전체적인 수요의 예측보다는 시간 단위 예측을 통해 자전거 재배치 작업의 효율성을 높이고자 하였다.

위와 같은 기계학습 방법의 장점 중 하나는 기존의 시계열 자료 분석의 한계였던 여러 가지 가정을 만족해야 한다는 점을 극복하고 탄력적으로 수요 예측을 할 수 있다는 것이다. 선행 연구로 김수현 등(2017)은 기계학습 방법 중 딥러닝(deep learning) 알고리즘인 순환신경망(recurrent neural network)을 이용해 시계열 자료를 분석하였다. 최민정 (2016)은 시계열 자료를 딥러닝 알고리즘 중 하나인 심층신경망(deep neural network)을 사용하여 제주도를 방문하는 관광객 수를 여행 목적별로 예측하였고, 예측정확도를 확인하기 위하여 기존에 사용되던 계절형 ARIMA 모형과 심층신경망 모형을 비교한 결과 향상되거나 비슷한 결과를 얻었다.

장다슬 (2016)은 시계열 자료 분석 시에 정상성 가정에 대한 단점의 한계를 보완하고자 비정상성 시계열 자료에 기계학습 방법 중 하나인 서포트 벡터 회귀 모형(support vector regression)을 사용해 서울지역의 가뭄을 예측을 하였다. 계절성과 과거시점의 변수를 이용해 모형을 구축하였고, 비정상성 시계열 자료에서 서포트 벡터 회귀모형에서 계절성은 예측정확도에 큰 차이를 주지는 않으나, 과거시점을 고려하면 더 높은 예측정확도를 얻을 수 있다고 하였다.

본 논문에서는 시계열 자료의 특성을 가지고 있는 공공 자전거 대여이력 자

료를 이용하여 전체 공공 자전거 대여소의 대여량의 합계를 기반으로 예측하는 방법이 아닌 시계열 군집분석(time series clustering)을 사용해 공공 자전거의 대여소를 군집으로 나누고, 각각의 군집 별로 기계 학습 방법인 서포트 벡터 회귀모형을 이용해 전체적인 서울시의 공공자전거 일일 수요량을 예측을 하고자 한다. 시계열 형태의 자료를 사용하기 때문에 기존의 군집분석방법에서 발전된 시계열 군집분석 방법을 사용 하였다.

시계열 군집분석 방법을 사용해 수요량을 예측한 경우로 Sohn 등(2016)은 기업의 전력 사용량의 패턴을 이용해 자료를 시계열 군집분석 방법으로 군집을 나누었고, 장기적 특성을 고려한 시계열 모형(fractional ARIMA model), 두 개의 계절 주기를 가지는 DSHW 모형(double seasonal Holt-Winter)을 사용하여 전력 사용량을 예측하였는데, 군집으로 나눈 뒤에 수요 예측모형을 만들어 정확한 수요 예측을 한 사례이다.

본 논문에서 시계열 군집분석을 함으로써 얻을 수 있는 장점은 장기수요예측을 할 때로 볼 수 있다. 장기수요예측이라고 함은 주로 미래의 수요에 대한 시설투자계획 수립을 위해 수행된다. 또한 예측의 정확성 여부에 따라 과잉투자를 방지하거나 또는 유발할 수 있으므로 합리적인 방법에 의해 예측의 정확성을 기하여야 한다(김성현, 2015). 기존의 서울시 전체 공공 자전거 수요 예측 분석은 전체 대여소의 모든 수요량의 합계를 사용하여 예측을 하였다. 이에 비해 시계열 군집분석 방법을 기반으로 하면 군집별로 최적의 모형을 만들 수 있어 기존보다 정확한 수요예측을 할 수 있는 장점이 있다. 그리고 각각의 대여소의 단기수요예측을 하는 경우에도 각각의 수백 개의 대여소별로 모형을 만드는 경우 시간적, 인적 자원의 효율이 저하된다. 하지만 시계열 군집분석을 이용해 특징을 가진 군집으로 묶어서 관리를 할 경우 보다 효율적으로 자원을 사용할 수 있는 장점이 있다.

## 2. 연구의 구성

본 논문은 총 5장으로 구성하였다. 1장에서는 연구의 배경 및 목적에 대해서 설명하고, 2장에서는 시계열 군집분석 방법과 서포트 벡터 회귀모형의 이론에 대해 설명하였다. 3장에서는 실제 분석에 사용할 자료에 대해 소개하면서 자료의 전처리 방법에 대해 설명하였고, 4장에서는 분석 결과를 평가한 뒤 비교하여 5장에서 연구의 결론과 향후 연구과제에 대해 설명해 마무리 하였다.

# 2. 분석 방법론

2장에서는 본 논문에서 사용된 분석들의 이론적인 배경과 방법에 대해 설명하였다. 2.1장에서는 시계열 군집분석의 방법에 대해 설명하고, 2.2장에서는 서포트 벡터 회귀모형의 방법에 대해 설명하였다.

## 2.1 시계열 군집분석

시계열 군집분석은 기존의 군집분석에서 파생된 군집분석 방법이다. Montero 와 Vilar (2014)가 제안한 방법으로 모형을 만들지 않고 시계열 자료원 변수의 패턴을 이용해 군집분석을 할 수 있는 방법과 모형을 만든 후 모수를 이용한 군집분석을 할 수 있는 방법, 그리고 예측 시점을 이용한 군집분석을 할 수 있는 방법이 있다. 그리고 이를 기반으로 TSclust R package(https://cran.r-project.org/web/packages/TSclust/TSclust.pdf)를 제안하였다. 시계열 군집분석은 매칭이 가능한 모든 두 시계열 간의 거리를 최소화 시키는 방법으로 집단 간의 유사성을 기반으로 군집화 하는 방법이다. 식 (2.1)은 두 시계열의 매칭을 위한 정의이다.

$$X_T = (X_1, X_2, \cdots, X_T)^\top,\ Y_T = (Y_1, Y_2, \cdots, Y_T)^\top, \qquad (2.1)$$
$$X = \{X_t, t \in \mathbb{Z}\},\ \ Y = \{Y_t, t \in \mathbb{Z}\}.$$

$M$ 이 관측값 간의 가능한 모든 매칭의 형태라고 할 경우 식 (2.2)와 같은 형태로 나타내고 정의할 수 있다.

$$r = ((X_{a1}, Y_{b1}),(X_{a2}, Y_{b2}), \cdots ,(X_{am}, Y_{bm})), \qquad (2.2)$$

$$\text{with } ai, bi \in 1,2,\cdots,T$$

$$\text{such that } a_1 = b_1 = 1, a_m = b_m = T.$$

### 2.1.1 시계열 군집분석의 거리함수

2.1에서 소개한 두 시계열 간의 거리를 계산하는 방법들을 거리함수(distance function)라고 한다. 비슷한 거리를 가지고 있는 시계열 자료의 쌍이라면 거리함수를 이용해 계산하면 낮은 거리함수 값을 가진다. 본 논문에서는 모형에 영향을 받지 않는 거리함수를 사용해 시계열 군집분석을 실시하였다. 다음 식 (2.3)은 본 연구에서 사용한 거리함수인 Minkowski 거리의 정의이다.

$$d_{Lq}(X_T, Y_T) = \left( \sum_{t=1}^{T} (X_t - Y_t)^q \right)^{1/q}. \qquad (2.3)$$

Minkowski 거리는 일반적으로 $q$ 에 따라 다른 두 가지 거리함수로 함께 사용 할 수 있는데, $L_q$-norm 거리로 정의할 수 있다. $T$는 식 (2.2)에서 정의를 따른다. $q=2$ 인 경우 $L_2$-norm 거리 혹은 Euclidean 거리라고 정의한다. 그리고 $q=1$ 인 경우는 $L_1$-norm 거리 혹은 Manhattan 거리라고 정의한다. Manhattan 거리의 경우 $X_t$와 $Y_t$의 차 들의 합으로 나타 낼 수 있고, Euclidean 거리의 경우는 $X_t$와 $Y_t$의 차 를  제곱한 값들을 합하고 제곱근을 취한 값이 된다. 즉, $X_t$와 $Y_t$의 가장 가까운 거리를 계산하는 것을 의미한다.

그 다음 사용한 거리함수는  DTW(Dynamic Time Warping) 거리를 사용하였다.  Berndt 과 Clifford (1994)에 제안한 거리함수로써, 동적시간와핑 거리라고 부를 수 있으며, 모형의 제약이 없이 접근할 수 있는 시계열 군집분석 거리함수의 대표적인 거리함수 이다. 동적시간와핑 거리는 두 시계열 $X_{ai}$와

$Y_{ai}$간의 거리를 최소화하는 방향으로 움직이며 쌍을 이루는 방법으로 누적 거리를 계산해 최소한의 누적 거리를 이루는 집단들을 군집화 시키는 거리 함수이다. 두 시계열 간의 근접성을 측정하는 방법으로, 계산한 값의 근접도에 기반 하여 사용하며 식 (2.4)로 정의 한다.

$$d_{DTW}(X_T, Y_T) = \min_{r \in M} \left( \sum_{i=1,2,\cdots,m} |X_{ai} - Y_{bi}| \right). \quad (2.4)$$

마지막으로 사용한 거리함수는 Caiado 등 (2006)이 제안한 Periodogram 거리이다. Periodogram 기반의 거리함수이고, 주기에 기반한 방법으로 다음 식 (2.5)처럼 $X_T$와 $Y_T$의 Peridogoram을 정의 할 수 있다.

$$I_{X_T}(\lambda_k) = T^{-1} \left| \sum_{t=1}^{T} X_t e^{-i\lambda_k t} \right|^2, \quad (2.5)$$
$$I_{Y_T}(\lambda_k) = T^{-1} \left| \sum_{t=1}^{T} Y_t e^{-i\lambda_k t} \right|^2.$$

이 때, 진동수는 $\lambda_k = 2\pi k/T$, $k=1,2,\cdots,n$ 과 $n=[T-1/2]$로 나타낼 수 있고, 식 (2.6)으로 거리함수를 정의할 수 있다.

$$d_P(X_T, Y_T) = \frac{1}{n} \sqrt{\sum_{k=1}^{n} (I_{X_T}(\lambda_k) - I_{Y_t}(\lambda_k))^2}. \quad (2.6)$$

값의 크기보다 상관 구조에 더 관심이 있을 경우에는 정규화 Periodogram 좌표 사이의 Euclidean 거리를 계산하면 향상된 결과를 얻을 수 있고, 식(2.7)

과 같이 정의 할 수 있다.

$$d_{NP}(X_T, Y_T) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(NI_{X_T}(\lambda_k) - NI_{Y_t}(\lambda_k))^2}, \qquad (2.7)$$

여기서, $d_{NP}$는 정규화(normalized)된 Periodogram 거리를 의미하고, 여기서 $N$은 정규화를 의미한다.

## 2.2 서포트 벡터 회귀

서포트 벡터 회귀는 임의의 실수값을 예측할 수 있도록 서포트 벡터 기계(support vector machine)의 훈련 자료를 클래스(class)로 구분하는 것에 사용한 방법을 확장하고 일반화 한 방법이다(Vanpnik 등, 1997). 서포트 벡터 기계 모형은 입력공간과 관련된 비선형(non-linear)문제를 고차원 공간의 선형문제로 대응시키는 방법이므로 수학적으로 분석하는 것이 용이한 장점을 가지고 있다(Hearst 등, 1998). 또한 조절해야하는 모수가 복잡하지 않고 다양하지 않아 간단하게 학습에 미치는 모수들을 설명할 수 있음에 비해 뛰어난 학습 성능을 보이는 장점을 가지고 있다.

서포트 벡터 회귀모형은 훈련 자료가 $\{(x_i, y_i), \cdots, (x_n, y_n)\}$ 이라면, $x \in R^n$ 은 표본에 대한 독립변수의 공간이고, $y \in R$ 은 $i = 1,2,\cdots,N$ 에 대응되는 목표 변수의 값을 의미한다. 이 때, 서포트 벡터 회귀의 목표는 실제 값 $y_i$으로부터 $\epsilon$(intensive parameter)이 최대가 되는 오차에 존재하면서, 가장 작은 기울기인 $w$를 가지는 함수를 의미하고 식 (2.8)으로 나타낼 수 있다.

$$f(x) = < w, x > + b,\ w \in R^n,\ b \in R, \qquad (2.8)$$

여기서, $b$는 절편을 의미하고, $< w, x >$는 $w$와 $x$의 내적을 의미한다. 다음으로 $x$ 값들의 회귀위험을 최소화 하는 가장 작은 $w$를 찾기 위해서는 볼록 최적화 문제(convex optimization problem)을 고려해야 하는 경우는 식 (2.9)를 통해 알 수 있다.

$$\text{minimize } \frac{1}{2}\|w\|^2, \tag{2.9}$$

$$\text{subject to } \begin{cases} y_i - < w, x_i > - b \le \epsilon \\ < w, x_i > + b - y_i \le \epsilon. \end{cases}$$

하지만 모든 학습데이터가 식 (2.9)를 만족할 수 없을 경우에는 일정 오차를 허용하는 여유 변수(slack variable)인 $\xi_i, \xi_i^*$를 이용해 식 (2.10)과 같은 최적화 문제를 다시 고려 할 수 있다.

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i, \xi_i^*), \tag{2.10}$$

$$\text{subject to } \begin{cases} y_i - < w, x_i > - b \le \epsilon + \xi_i \\ < w, x_i > + b - y_i \le \epsilon + \xi_i \end{cases}$$

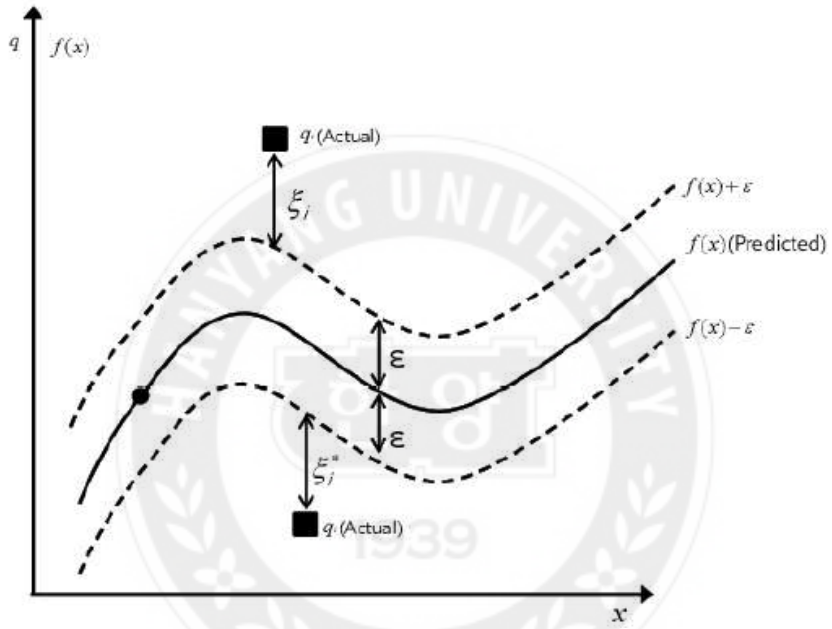
$$\xi_i, \xi_i^* \ge 0.$$

여기서 $C$ 는 모형의 복잡도를 결정하는 모수로 추정오차의 패널티(penalty)를 의미한다. $C$ 가 커지면 서포트 벡터 회귀모형은 오차를 엄격하게 규제하게 되고, 반대로 $C$ 가 0으로 작아지면 오차를 보다 크게 허용하고, 단순한 모형이 된다. 적절한 $C$ 의 선택은 서포트 벡터 회귀모형에서 중요한 문제로 성능의 결과를 달라지게 할 수 있다(Smola 등, 1998).

서포트 벡터 회귀모형에서는 일반화 능력을 최대화하는 모형을 만들고, 오차를 구하기 위해 $\epsilon$-무감각 손실 함수를 사용한다. $\epsilon$-무감각 손실 함수는 서포트 벡터 회귀 식인 식 (2.8)의 오차가 $\pm\epsilon$의 범위에 있으면 범위 안의 오차를 무시한다. $\epsilon$-무감각 손실 함수는 식 (2.11)로 나타낼 수 있고, [그림 2.2]로

표현 할 수 있다(Lu 등, 2009).

$$L_{\epsilon}(y, f(x)) = (|y - f(x)| - \epsilon)_{+} = \begin{cases} 0 & \text{if } |y - f(x)| \le \epsilon \\ |y - f(x)| - \epsilon, & \text{if } |y - f(x)| > \epsilon . \end{cases} \quad (2.11)$$




[그림 2.1] 서포트 벡터 회귀모형의 $\epsilon$-무감각 손실함수

식 (2.10)의 볼록 최적화 문제는 라그랑지 승수를 도입하면 이차계획(quadratic programming)의 최소화를 통해 나타낼 수 있다. 식 (2.10)을 라그랑지 함수로 나타내면 식 (2.12)과 같다.

$$L_p = \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\eta_i \xi_i + \eta_i^* \xi_i^*) \qquad (2.12)$$
$$- \sum_{i=1}^{n} \alpha_i (\epsilon + \xi_i - y_i + < w, x_i > + b)$$
$$- \sum_{i=1}^{n} \alpha_i^* (\epsilon + \xi_i^* - y_i + < w, x_i > - b) .$$

식 (2.12)에서 $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$는 라그랑지 승수를 의미하고, 이는 0보다 크거나 같아야 한다. 식 (2.12)을 최적화하기 위해 $w, b, \xi_i, \xi_i^*$ 에 대하여 편미분을 한 하면 식 (2.13)로 나타낼 수 있다.

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0, \qquad (2.13)$$
$$\frac{\partial L_p}{\partial b} = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0,$$
$$\frac{\partial L_p}{\partial \xi_i} = C - \alpha_i - \eta_i = 0,$$
$$\frac{\partial L_p}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0.$$

식 (2.13)의 편미분 식들을 식 (2.12)에 대입해 $\alpha_i, \alpha_i^*$ 를 최대화하면 쌍대 최적화 문제를 해결할 수 있다. 식 (2.14)은 이를 정의한 식이다.

$$\text{maximize} \quad L_p = -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)<x_i,x_j> \\ -\epsilon\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*), \tag{2.14}$$

$$\text{subject to } \sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0 \text{ and } \alpha_i, \alpha_i^* \in [0,C] .$$

라그랑지 승수 $\alpha_i, \alpha_i^*$ 중 0이 아닌 값이 회귀 계수를 추정하는데 사용된다. 이 때의 값을 서포트 벡터라고 한다. 식 (2.13)에서 $w=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)x_i$ 이므로 서포트 벡터 회귀 추정함수는 식 (2.15)로 나타낼 수 있다.

$$\begin{aligned} f(x) &= <w,x>+b \\ &= \sum_{i=1}^{N}(\alpha_i-\alpha_i^*)<x_i,x>+b . \end{aligned} \tag{2.15}$$

$b$는 Karush-Kuhn_Tucker(KKT) 조건을 이용하면 계산할 수 있다. KKT 조건은 최적해에서 쌍대변수와 제약식의 곱이 0이 되는 조건이다. 이는 식 (2.16)로 표현 할 수 있다.

$$\begin{aligned} &\alpha_i(\epsilon+\xi_i-y_i+<w,x_i>+b)=0, \\ &\alpha_i^*(\epsilon+\xi_i^*+y_i-<w,x_i>-b)=0, \\ &(C-\alpha_i)\xi_i=0, \\ &(C-\alpha_i^*)\xi_i^*=0. \end{aligned} \tag{2.16}$$

식 (2.16)의 조건을 통해 $b$를 구할 수 있고, 식 (2.17)로 나타낼 수 있다.

$$\begin{aligned}\max\{-\epsilon+y_i-<w,x_i>\mid \alpha_i<C \text{ or } \alpha_i^*>0\} \\ \leq\ b\ \leq\ \min\{-\epsilon+y_i-<w,x_i>\mid \alpha_i>C \text{ or } \alpha_i^*<0\}\end{aligned} \tag{2.17}$$

식 (2.17)의 과정을 통해 서포트 벡터 회귀모형을 구축할 수 있다. 실제 자료에서 자주 사용하게 되는 비선형 서포트 벡터 회귀는 커널을 통해 모형을 구축할 수 있다. 비선형 사상함수 $\Phi$를 이용해 입력공간의 독립변수를 고차원의 특성공간으로 사상시켜 비선형 서포트 벡터 회귀모형을 만들 수 있다. $\Phi$는 입력 공간의 독립변수를 고차원의 특성공간으로 반영하는 변환함수 이다(최희령, 2015). 다음 식 (2.18)은 비선형 서포트 벡터 회귀모형식이다.

$$f(x)=<b\bullet\Phi(x)>+b. \tag{2.18}$$

라그랑지 쌍대 최적화 문제는 변환함수 $\Phi$ 를 통해 다음 식 (2.19)로 나타낼 수 있다.

$$\begin{aligned}\text{maximize}\quad L_p=&-\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)<\Phi(x_i),\Phi(x_j)> \\ &-\epsilon\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*),\end{aligned} \tag{2.19}$$

$$\text{subject to}\ \sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0\ \text{ and }\ \alpha_i,\alpha_i^*\in[0,C].$$

자료의 분포에 따라 각각 알맞은 다른 커널 함수를 사용할 수 있는데, 커널 함수는 식 (2.20)으로 나타낼 수 있다.

$$K(x_i, x_j) = < \Phi(x_i), \Phi(x_j) > . \quad (2.20)$$

식 (2.13)에서와 마찬가지로 $w = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \Phi(x_i)$이고, 비선형 서포트 벡터 회귀모형식은 식 (2.21)으로 나타낼 수 있다.

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K < x_i, x > + b. \quad (2.21)$$

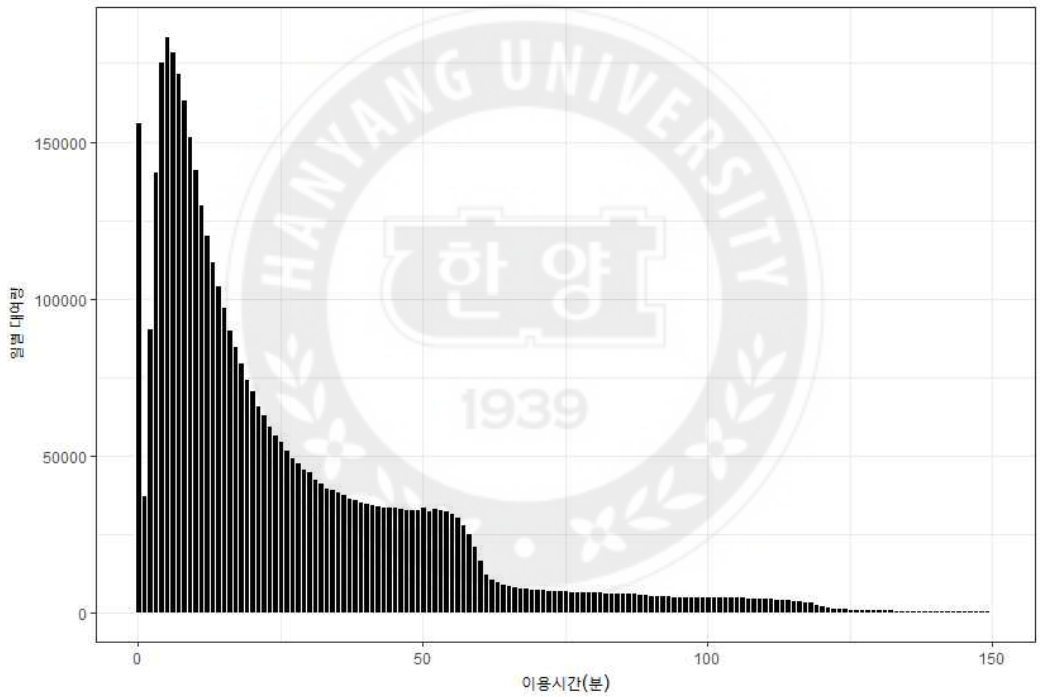
# 3. 실제 자료 분석

3장에서는 분석에 쓰인 실제 자료에 대해 다루고, 분석에 쓰인 방법에 대해 설명하였다. 3.1장에서는 자료의 소개와 전처리에 대하여 설명하였고, 3.2장에서는 분석용 자료의 변수와 구성에 대하여 설명하였다. 3.3장에서는 분석의 흐름과 본 분석에서 사용한 두 모형의 비교 방법에 대해 설명 하였다.

## 3.1 자료 소개 및 전처리

본 논문의 분석에 쓰인 자료는 서울시설공단에서 관리하고 있는 원시 대여이력 형태의 자료이다. 자료에 등록된 대여소들은 2016년 9월까지 확충한 대여소를 포함, 서울시의 모든 공공자전거 대여소의 대여이력을 모은 자료이다. 여기서 대여이력이란 한 건의 대여와 반납으로 이루어진 사건을 말하며, 한 건의 대여이력이 하나의 행으로 기록되어 있다. 자료의 기간은 2016년 9월부터 2017년 9월까지 총 395일로 이루어져 있다. 기간 동안의 자료의 양은 455만 5311건으로 나타났다. 원시 자료에서 사용한 변수는 대여소, 대여일시, 그리고 이용시간 변수이다. 분석에 앞서 정확한 수요를 예측하기 위해여 결측값(missing value)과 이상값(outlier)을 처리하고, 비정상적인 이용형태의 대여이력을 제거하는 자료의 전처리 과정을 시행하였다. 그리고 다시 대여소별로 일별 대여량을 구하여 원시 자료를 분석에 맞는 자료 형태로의 전처리 과정을 하였다. 먼저 비정상적인 대여이력에 대한 처리를 위해 자료의 범위를 확인한

결과 최소 0분에서 최대 5400분까지로 자료의 범위가 매우 넓었다. [그림 3.1]은 이용시간에 따른 대여이력의 분포를 보기 위한 그림이다. 모든 범위를 그림으로 나타내지 않고, 가장 많은 자료가 몰려있는 부분을 확대하였다. 특이한 비정상적 이용형태인 이용시간이 0분인 경우의 빈도수를 확인할 수 있고, 이용시간이 일정 시간 이후로는 자료에 영향을 줄 수 없는 매우 적은 빈도수를 가지고 있었다.




[그림 3.1] 이용시간에 따른 대여이력의 분포

[그림 3.1]에서 첫 번째 막대가 이용시간이 0분인 경우에 대한 대여 이력이다. 자료의 기간인 395일 동안 이용시간이 0분인 총 15만 6129건의 비정상적인 이용형태가 나타났다. 이와 같은 현상에 대해 대전시 공공 자전거 시스템의 시각화를 연구한 문현수 등(2016)은 비정상적인 대여이력의 이유를 자전거

단말기 시스템의 오류, 자전거의 고장, 또는 이용자의 변심으로 해석하였다.

다음으로 일반적인 비정상적인 이용형태인 이상값(outlier)를 해결하기 위해 상자그림(Box plot)을 그렸다. [그림 3.1]의 막대그래프를 확인해보면 오른쪽으로 꼬리가 굉장히 긴 자료의 형태를 가지고 있기 때문에 이상값이 존재할 확률이 크다. 일반적인 이상값의 하한선은 제 1사분위수에서 사분위범위의 1.5의 곱을 빼는 방법을 사용하지만, 본 자료에서는 그 값이 음수로 나오기 때문에 이상값의 하한은 정하지 않았다. 자료의 특징상 이상값으로 판단할 수 있는 이용시간이 음수인 형태의 자료는 발견할 수 없었다. 다음으로 이상값의 상한을 구하는 형태는 식 (3.1.1)과 같다.

$$Q3 + 1.5\,IQR, \qquad (3.1.1)$$

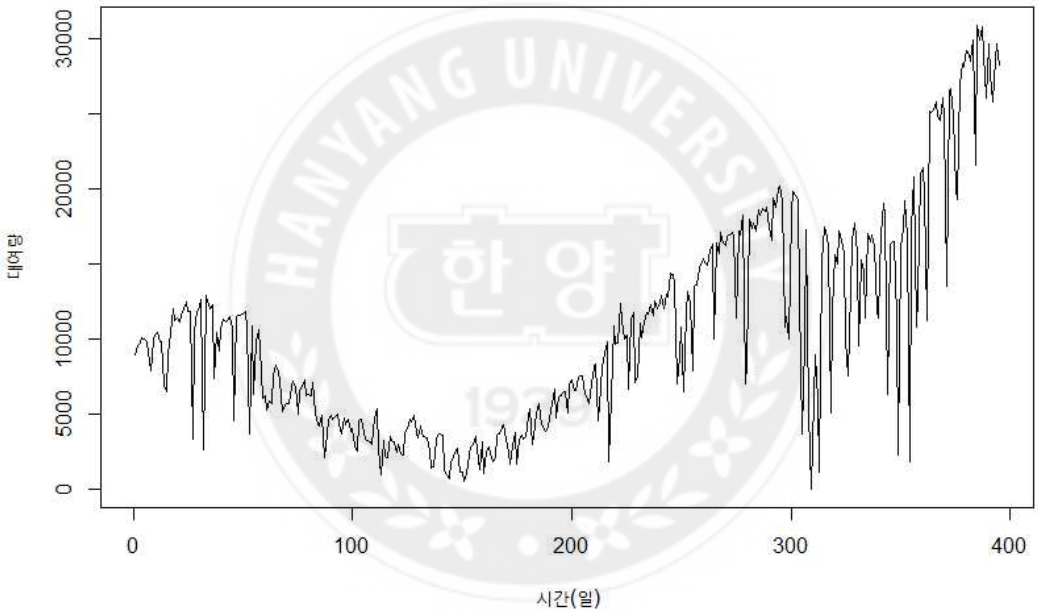
여기서 $Q3$는 제3사분위수를 말하고, $IQR$은 사분위범위(Interquartile range)로 $Q3$인 제3사분위수에서 $Q1$인 제1사분위수 간의 거리를 말한다. 자료의 전처리를 끝내고 남은 자료의 양은 417만 1850건 이다. 총 38만 3461건의 자료가 비정상적인 형태의 자료로 확인되어 제외되었다. [표 3.1]은 전처리를 마친 자료의 기초통계량을 구한 표이다.

[표 3.1] 비정상적인 대여이력을 제외한 자료의 기초통계량

| 최솟값 | 제1사분위수 | 중앙값 | 평균 | 제3사분위수 | 최댓값 |
|---|---|---|---|---|---|
| 1 | 8 | 16 | 23.36 | 35 | 88 |

이용시간의 최솟값은 1분이고, 제1사분위수는 8분, 중앙값은 16분, 평균은 23.36분, 제3사분위수는 35분, 최댓값은 88분으로 나타났다. [그림 3.3]은 자

료의 전처리를 마친 후 일별 대여이력을 모두 세어 최종적으로 자료는 총 395일의 일별 대여량 형태로 변환하였다. [그림 3.3]은 서울시 전체 대여소의 일별 대여량을 시계열 그림으로 나타낸 것이다. 시간의 변화인 계절에 따라 감소하였다가 점점 증가하는 추세성분을 가지고 있고, 자료들이 일정한 기간을 주기로 주기성을 가지는 계절성분을 가진 비정상성 시계열 자료이다. Y축은 일별 대여량이며, X축은 일 단위의 시간 단위이다.




[그림 3.2] 서울시 전체 대여소의 일별 대여량 시계열 그림

## 3.2 연구 자료의 구성

시계열 군집분석을 사용하지 않은 기존 모형과 시계열 군집분석을 기반으로 한 모형은 서로 다른 자료의 형태를 이용해 분석을 하였다.

### 3.2.1 시계열 군집분석을 사용하지 않은 기존 모형

먼저 시계열 군집분석을 사용하지 않은 서포트 벡터 회귀모형에 사용하는 연구 자료를 구성하고자 하였다. 본 논문에서는 기상, 주변시설, 교통 연계여부 등의 다른 변수를 제외하고 전 시점의 자료를 이용해 서포트 벡터 회귀모형을 구성하였다. 예측 전 1주일, 2주일, 3주일의 일별 대여량을 독립 변수로 사용하였다. 전 1주일은 7일로 구성되어 있으므로 대여량을 변수로 사용하는 모형은 7개의 설명 변수를 가지고 있고, 전 2주일 대여량을 변수로 사용하는 모형은 총 14개의 변수를 설명변수로 가진다. 전 3주일 대여량을 변수로 사용하는 모형은 총 21개의 변수를 설명변수로 모형을 만들게 된다.

서포트 벡터 회귀모형은 훈련용 자료와 평가용 자료로 나누어 모형을 학습시키고, 모형을 평가를 할 수 있다. 이는 모형의 과적합 문제를 해결하기 위한 방법이다. 과적합 문제란 모형이 훈련용 자료에 맞추어져 훈련용 자료의 예측 정확도는 높으나 새로운 관측치나 새로운 자료를 이용해 예측할 때 예측도가 현저히 떨어지게 되는 문제이다. 훈련용 자료와 평가용 자료는 여러 비율을 기준으로 나누어 사용할 수 있다. 전 1주일, 2주일, 3주일 변수를 가지는 모형을 각각 다른 비율의 훈련용 자료와 평가용 자료로 나누어 분석에 사용하였다. 각각 70%와 30%, 80%와 20%, 90%와 10%로 총 3개의 비율로 나누어

3개의 각각 다른 전 시점의 변수들과 합쳐 총 9개의 분석용 자료를 만들었다. 해당 자료를 이용해 9개의 서포트 벡터 회귀 모형을 만들고, 각각의 모형마다 평균 제곱근 오차(root mean square error) 식(3.1)과 평균 절대오차(mean absolute error) 식 (3.2)을 구해 가장 낮은값을 가지고 있는 모형을 찾아 최종 모형으로 선택하고자 하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}. \quad (3.1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left| Y_i - \hat{Y}_i \right|. \quad (3.2)$$

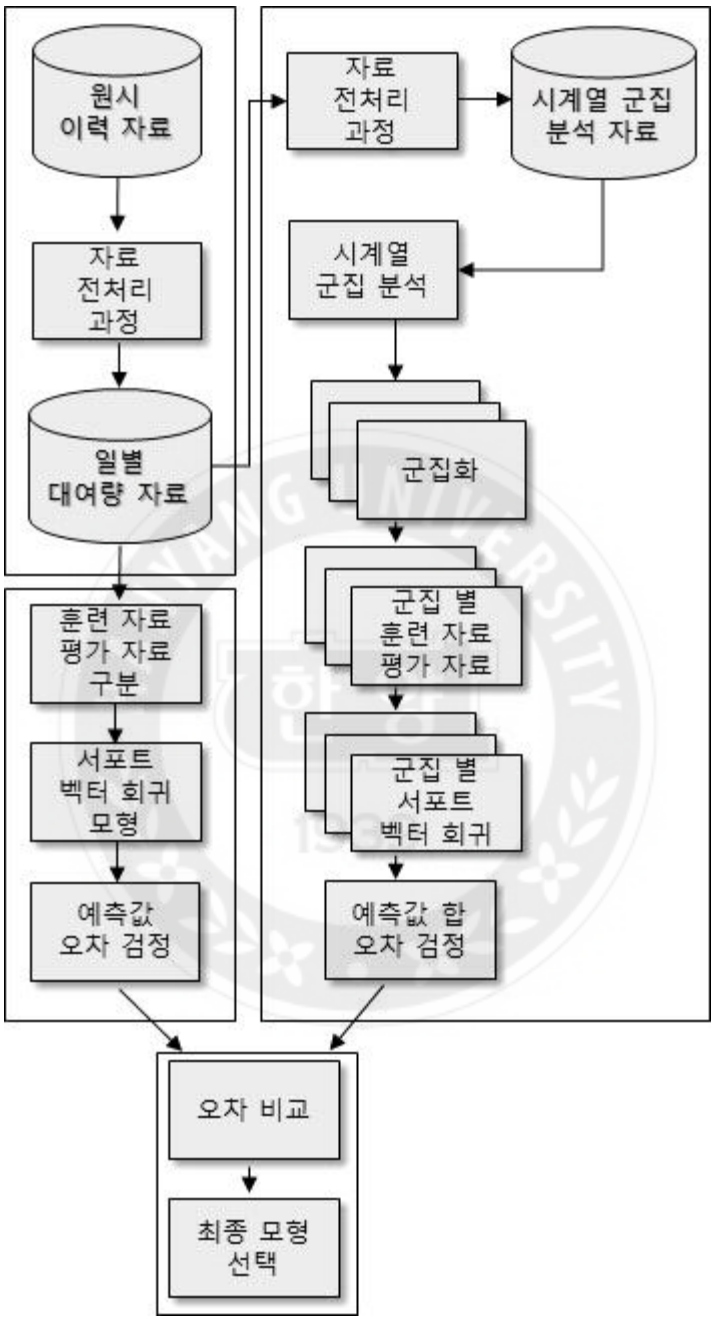
### 3.2.2 시계열 군집분석을 기반으로 한 모형

시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형은 먼저 시계열 군집분석을 실시한 후 각각의 군집마다 전 시점이 다른 1,2,3주의 변수를 이용해 서포트 벡터 회귀모형을 만들어 평균 제곱근 오차와 평균 절대 오차가 가장 낮은 모형을 선택하였다. 시계열 군집분석을 위해서는 새로운 자료의 형태로 일별 대여량을 정리하였다. 자료 형태는 시계열 군집분석을 위한 자료의 형태로 재구성하여, 하나의 행이 하나의 대여소의 정보를 의미하도록 하였다. 각 열은 1부터 439까지 총 439개의 대여소를 뜻한다. 대여소 이름 열은 각각의 대여소들의 이름이다. 그 뒤로는 1일부터 395일까지 총 395개의 열로 구성이 된다. 이는 전체 기간인 395일 각각의 일별 대여량을 의미한다. 마지막으로 군집열은 시계열 군집분석을 실시한 후 결과인 군집의 번호가 들어갈 공간으

로 구성되었다. 시계열 군집분석이 완료 된 후에는 각각의 군집별로 다른 변수를 사용해 최적의 서포트 벡터 회귀모형을 만들고 예측값을 합친 최종 예측값을 구해 시계열 군집분석을 사용하지 않은 모형과 평균 제곱근 오차와 평균 절대오차를 비교하였다.

## 3.3 예측 모형 구성 방법

본 논문의 목적은 시계열 군집분석을 기반으로 한 모형과 사용하지 않는 모형을 비교하는 것이다. 3.2.1장에서 만든 시계열 군집분석을 사용하지 않은 서포트 벡터 회귀모형과 같은 조건으로 3.2.2장에서 만든 시계열 군집분석을 사용한 서포트 벡터 회귀모형을 비교하는 것이 목적이다. 3.3장에서는 각각 3.2.1장과 3.2.2장에서 가장 잘 적합된 최적의 모형을 찾아서 비교하기 위한 연구의 흐름을 소개하고자 한다. 다음 [그림 3.4]는 예측 모형 구성 방법을 볼 수 있도록 정리한 흐름도 이다.




[그림 3.3] 수요예측 모형 구성 흐름도

# 4. 분석 결과 및 비교

4장 에서는 시계열 군집분석을 사용한 서포트 벡터 회귀모형과 사용하지 않은 두 모형의 결과를 비교하였다. 서포트 벡터 회귀모형은 커널함수와 커널에 따른 모수 그리고 손실함수의 모수에 따라 결과가 달라질 수 있다. 그렇기 때문에 각 모형을 모두 분석해보고 최적의 모형을 찾아 결과로 설명하였다. 4.1장에서는 시계열 군집분석을 사용하지 않은 모형의 결과를 설명하고, 4.2장에서는 시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형의 결과를 설명한다. 4.3장에서는 두 결과를 비교해 최종 모형을 선택하였다.

## 4.1 시계열 군집분석을 사용하지 않은 서포트 벡터 회귀모형

자료의 형태가 시계열의 형태를 가지고 있기 때문에 최적의 서포트 벡터 회귀모형을 만들기 위해서 자료에 적합한 커널함수를 찾아 분석을 진행하였다. 서포트 벡터 회귀모형 구축 시 관측된 자료에 대한 사전 정보가 없을 때는 RBF(radial basis function) 커널함수가 주로 사용된다(Karatzoglou 등 2006). 본 논문에 사용된 커널함수들은 총 4개로 Linear, RBF, Polynomial, Sigmoid를 사용하였고, 각 커널에 따른 모수를 변경하면서 최적의 모형을 찾았다. 손실함수의 모수 또한 각각 다른 모수를 사용해 최적의 모형에 맞는 모수를 찾았다. 최적의 결과를 얻어낸 서포트 벡터 회귀모형은 Linear 커널함수

를 사용하고 손실함수의 모수(cost parameter)가 0.1일 때 가장 좋은 결과를 얻을 수 있었다. 모형을 평가하기 위해서 실제값과 예측값의 차이를 평균 제곱근 오차와 평균 절대오차를 이용해 오차율을 구하였다. 3.2.1장에서 만든 총 9개의 연구 자료로 서포트 벡터 회귀모형을 만든 결과는 [표 4.1]과 같다.

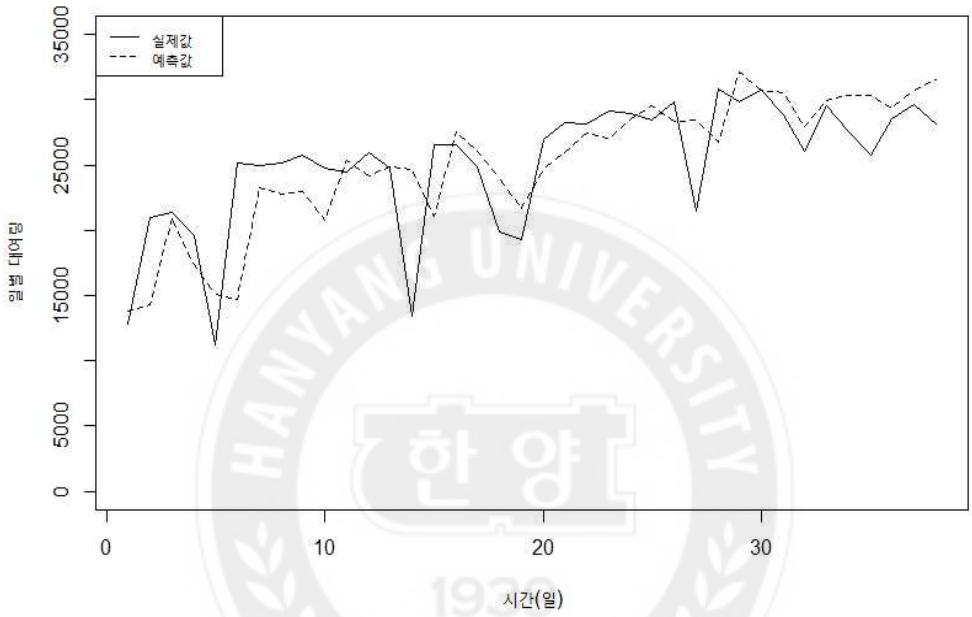
[표 4.1] 시계열 군집분석을 사용하지 않은 서포트 벡터 회귀모형의 결과

| 모형 | RMSE | MAE |
|---|---|---|
| 훈련용 70%, 전 1주일 대여량이 변수인 모형 | 4835 | 3549 |
| 훈련용 80%, 전 1주일 대여량이 변수인 모형 | 4775 | 3443 |
| 훈련용 90%, 전 1주일 대여량이 변수인 모형 | 4367 | 3199 |
| 훈련용 70%, 전 2주일 대여량이 변수인 모형 | 4741 | 3426 |
| 훈련용 80%, 전 2주일 대여량이 변수인 모형 | 4418 | 3130 |
| 훈련용 90%, 전 2주일 대여량이 변수인 모형 | **3728** | **2726** |
| 훈련용 70%, 전 3주일 대여량이 변수인 모형 | 4767 | 3511 |
| 훈련용 80%, 전 3주일 대여량이 변수인 모형 | 4460 | 3194 |
| 훈련용 90%, 전 3주일 대여량이 변수인 모형 | 3906 | 2933 |

RMSE: 평균 제곱근 오차, MAE: 평균  절대오차

가장 좋은 모형은 훈련용 자료 90%를 사용하고 변수는 전 2주일 일별 대여량 14개를 변수로 사용한 모형이며, 평균 제곱근 오차는 3728, 평균 절대오차는 2726으로 가장 낮은 값을 가지고 있어 최적의 모형으로 판단하였다. 훈련용 자료로 학습시킨 자료의 양이 많을 때 평균 제곱근 오차와 평균 절대오차

가 떨어지는 영향을 보였으나, 변수의 개수가 많아지는 경우는 크게 영향을 미치지 않았다. [그림 4.1]은 최적의 서포트 벡터 회귀모형을 이용해 실제값과 예측값을 시계열 그림으로 나타낸 것이다.




[그림 4.1] 시계열 군집분석을 사용하지 않은 서포트 벡터 회귀모형의 예측값 시계열 그림

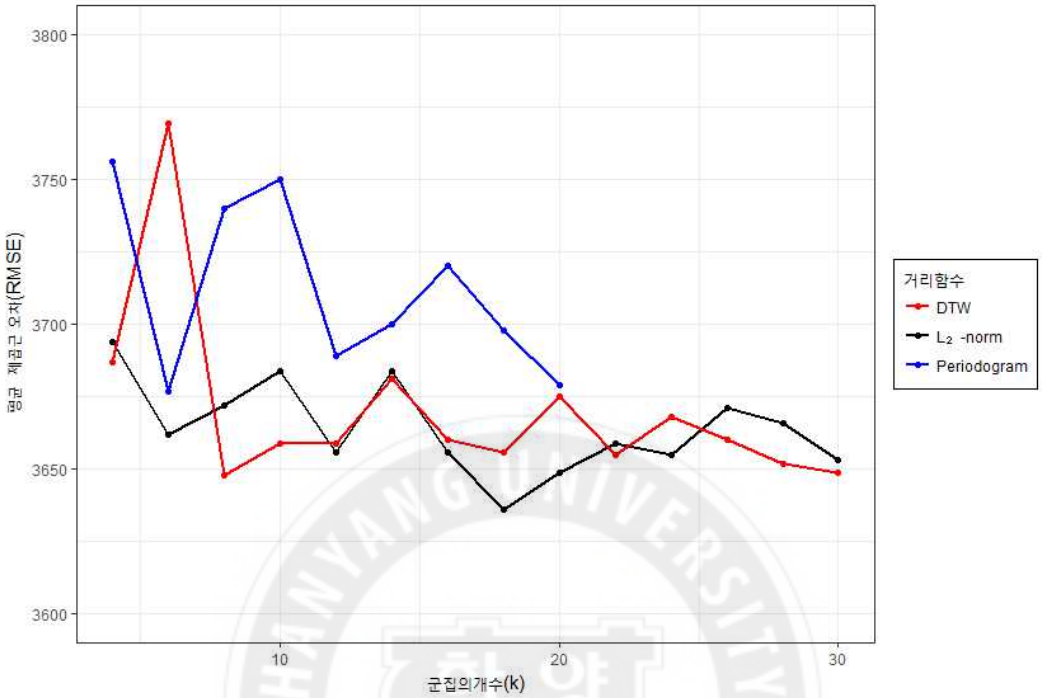
평가에 사용된 평가용 자료의 길이는 총 38일로 38개의 일별 대여량을 실제값으로 가지고 있다. 자세한 예측값과 실제값의 비교를 위해 95% 기준으로 실제값의 신뢰구간을 구하였고 예측값들이 신뢰구간 안에 존재하는지 확인하였다.

## 4.2 시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형

4.1장의 모형과는 달리 시계열 군집분석을 수행하면 군집들을 특성별로 파악할 수 있고 각각 군집별로 최적의 모형을 선택할 수 있기 때문에 더 정확한 예측모형을 구축하는 것이 가능해진다. 4.2.1장에서는 시계열 군집분석의 결과를 설명하고, 4.2.2장에서는 각 군집별 서포트 벡터 회귀모형의 결과에 대해 설명하였다.

### 4.2.1 시계열 군집분석의 결과

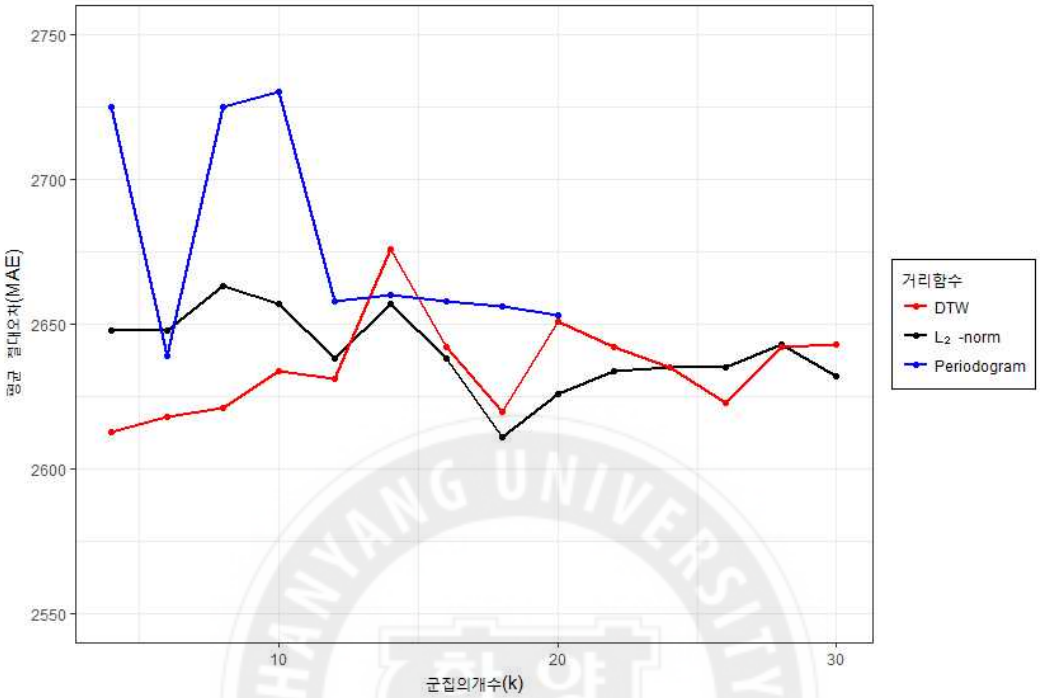
본 논문에서는 시계열 군집분석을 위해 $L_2$-nrom, DTW(Dynamic Time Warping), Periodogram 3가지의 거리함수를 사용하였다. 각각의 거리함수의 중심점 계산 방법은 정규화를 사용해 중심점을 계산하는 방법을 사용하였으며, 적절한 군집의 개수($k$)를 찾기 위해 군집의 개수($k=4,6,8,\cdots,30$)를 늘려가면서 분석을 진행하였다. 적절한 군집의 개수를 찾기 위해서 두 단계로 나누어 접근을 하였다. 먼저 군집의 개수가 달라질 때 마다 각 군집의 4.1장에서 선택한 서포트 벡터 회귀모형을 동일한 조건으로 만들어 평균 제곱근 오차와 평균 절대오차가 낮아 질 때의 군집의 개수를 적절한 군집의 개수로 정하는 격자 탐색방법(grid search)을 사용하였다. 그리고 첫 번째 단계에서 선택된 군집분석 모형을 다시 각 군집별 최적의 서포트 벡터 회귀모형을 만들어 비교를 한 후 최종 군집분석 모형을 선택하고자 한다. [그림 4.2]과 [그림 4.3]은 각 거리함수별 군집분석을 기반으로 한 서포트 벡터 회귀모형의 평균 제곱근 오차와 평균 절대오차를 꺾은선 그래프로 나타낸 것이다.




[그림 4.2] 각 거리함수별 군집분석의 평균 제곱근 오차 비교

DTW: 동적시간와핑 거리

먼저 [그림 4.2]는 거리함수별 군집분석의 평균 제곱근 오차를 비교한 꺾은선 그래프이다. Periodogram 거리함수를 사용한 군집분석 모형은 군집의 개수가 20개를 넘어 가면서 빈 군집이 만들어져 20개 이상의 군집은 의미가 없는 것으로 판단하였다. DTW 거리함수를 사용했을 때는 군집의 개수가 8개일 때 최적의 모형이라고 판단하였고, $L_2$-norm 거리함수를 사용했을 때는 군집의 개수가 18개 일 때 최적의 모형이라고 판단하였다.




[그림 4.3] 각 거리함수별 군집분석의 평균 절대오차 비교

DTW: 동적시간와핑 거리

[그림 4.3]은 거리함수별 군집분석의 평균 절대오차를 꺾은선 그래프로 비교한 것이다. 마찬가지로 Periodogram 거리함수를 사용한 군집분석 모형은 군집의 개수가 20개를 넘어 가면서 빈 군집이 만들어져 20개 이상의 군집은 의미가 없는 것으로 판단하였다. DTW 거리함수를 사용하였을 때 군집의 개수가 4개, 6개, 18개 일 때 낮은 평균 절대오차를 가지고 있어 최적의 모형이라고 판단하였다. $L_2$-norm 거리함수를 이용했을 때는 군집의 개수가 18개 일 때 최적의 군집분석 모형을 가진 것으로 나타났다.

### 4.2.2 군집별 서포트 벡터 회귀모형의 결과

다음 [표 4.2]는 [그림 4.2]와 [그림 4.3]을 합쳐 거리함수별 평균 제곱근 오차와 평균 절대오차를 숫자로 나타낸 표이다.

[표 4.2] 거리함수별 평균 제곱근 오차와 평균 절대오차의 비교

| 군집 개수 ($k$) | $L_2$-norm | | DTW | | Periodogram | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| 4 | 3694 | 2648 | **3687** | **2613** | 3756 | 2725 |
| 6 | 3662 | 2648 | **3769** | **2618** | 3677 | 2639 |
| 8 | 3672 | 2663 | 3648 | 2621 | 3740 | 2725 |
| 10 | 3684 | 2657 | 3659 | 2634 | 3750 | 2730 |
| 12 | 3656 | 2638 | 3659 | 2631 | 3689 | 2658 |
| 14 | 3684 | 2657 | 3681 | 2676 | 3700 | 2660 |
| 16 | 3656 | 2638 | 3660 | 2642 | 3720 | 2658 |
| 18 | **3636** | **2611** | **3656** | **2620** | 3698 | 2656 |
| 20 | 3649 | 2626 | 3675 | 2651 | 3679 | 2653 |
| 22 | 3659 | 2634 | 3655 | 2642 | | |
| 24 | 3655 | 2635 | 3668 | 2635 | | |
| 26 | 3671 | 2635 | 3660 | 2623 | | |
| 28 | 3666 | 2643 | 3652 | 2642 | | |
| 30 | 3653 | 2632 | 3649 | 2643 | | |

RMSE: 평균 제곱근 오차, MAE: 평균 절대오차, DTW: 동적시간와핑 거리

각 거리함수별 평균 제곱근 오차와 평균 절대오차가 낮은값을 기준으로 종합적으로 비교해 보면 $L_2$-norm 거리함수를 사용하고 군집의 개수가 18개 일

때, DTW 거리함수를 사용하고 군집의 개수가 4개, 6개, 18개 일 때 최적의 군집분석 모형을 가지고 있다고 판단하였다. [표 4.3]은 다음 단계로 선택된 4개의 군집분석 모형을 통해서 각각의 군집별로 서포트 벡터 회귀모형을 만들어 군집마다 최적의 모형을 찾아 예측값을 구한 후 합계를 비교한 표이다.

[표 4.3] 최종 군집분석 모형 선택을 위한 비교

| 모형 이름 | RMSE | MAE |
|---|---|---|
| $L_2$-norm, (k=18) | **3252** | **2399** |
| DTW, (k=4) | 3526 | 2518 |
| DTW, (k=6) | 3382 | 2432 |
| DTW, (k=18) | 3387 | 2473 |

RMSE: 평균 제곱근 오차, MAE: 평균 절대오차, DTW: 동적시간와핑 거리, k:군집수

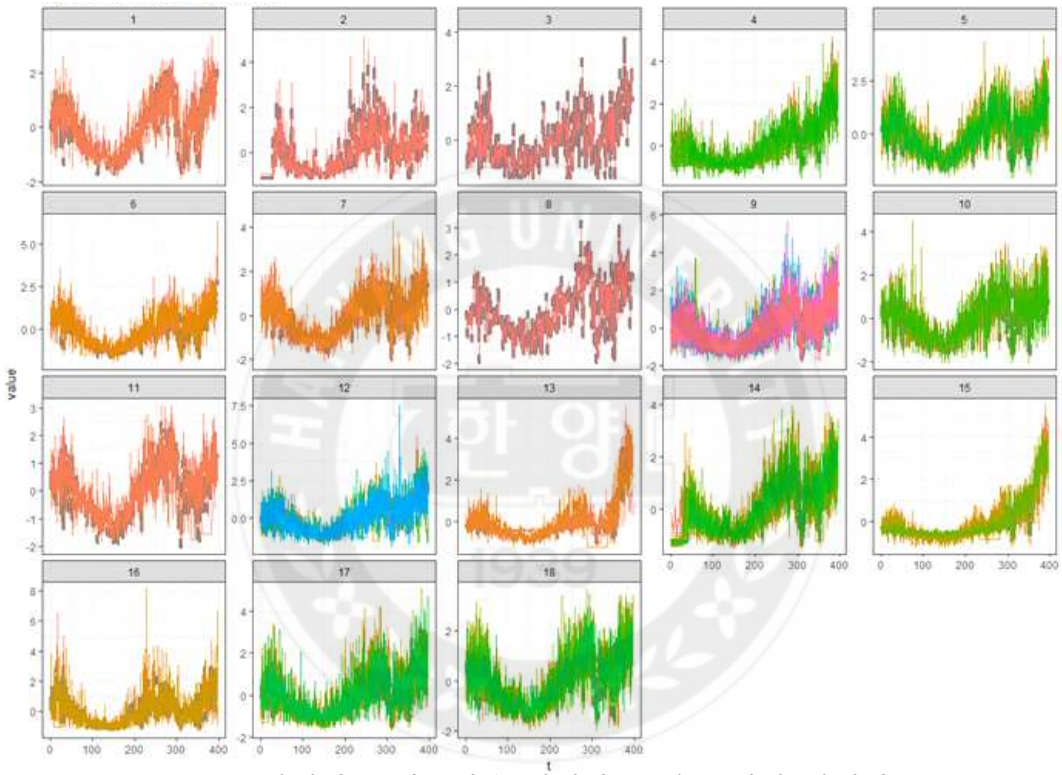
[표 4.3]에서 가장 낮은 평균 제곱근 오차와 평균 절대오차를 가진 모형은 $L_2$-norm, (k=18)으로 확인할 수 있다. 평균 제곱근 오차는 3252, 평균 절대오차는 2399로 가장 낮은값을 가지고 있는 모형이다. 다음 [표 4.4]는 $L_2$-norm, (k=18)모형의 군집별 최적의 서포트 벡터 모형을 찾기 위해서 변수의 개수별로 각각 평균 절대오차를 구한 표이다.

[표 4.4] $L_2$-norm, (k=18) 모형의 군집별 최적의 서포트 벡터 회귀모형의 평균 절대 오차

| 군집 이름 | 전 1주일 변수 | 전 2주일 변수 | 전 3주일 변수 |
|---|---|---|---|
| 군집 1 | 88.811 | **73.953** | 76.705 |
| 군집 2 | 130.361 | **115.117** | 117.884 |
| 군집 3 | 2.207 | 2.121 | **2.0744** |
| 군집 4 | 557.558 | **488.630** | 527.954 |
| 군집 5 | 87.731 | **71.142** | 78.069 |
| 군집 6 | 82.854 | **65.207** | 67.496 |
| 군집 7 | 8.725 | **7.546** | 7.660 |
| 군집 8 | 4.509 | **3.827** | 3.951 |
| 군집 9 | 253.604 | **219.150** | 237.566 |
| 군집 10 | 7.165 | 6.397 | **6.050** |
| 군집 11 | 14.020 | **10.916** | 11.588 |
| 군집 12 | 186.812 | 142.395 | **140.123** |
| 군집 13 | 422.251 | 375.566 | **361.556** |
| 군집 14 | 146.783 | **123.228** | 133.983 |
| 군집 15 | 1252.049 | **1099.425** | 1172.058 |
| 군집 16 | 81.541 | **62.568** | 68.721 |
| 군집 17 | 71.889 | **63.402** | 68.457 |
| 군집 18 | 131.406 | **101.322** | 107.375 |

[표 4.4]에서는 정확한 구분을 위해서 소수점 3번째 자리까지의 값을 사용하였다. 군집 1,2,4,5,6,7,8,9,11,14,15,16,17,18 은 전 2주일 일별 대여량을 변수로 한 서포트 벡터 회귀모형에서 가장 낮은 평균 절대오차를 보여 해당 모형으로 예측을 하였고, 군집 3,10,12,13은 전 3주일 일별 대여량을 변수로

한 서포트 벡터 회귀모형에서 가장 낮은 평균 절대오차를 보여 해당 모형으로 예측을 하였다. 그리고 시계열 군집분석을 사용하지 않은 모델과 비교를 할 모형으로 $L_2$-norm, (k=18)모형을 최종 선택하였다.




[그림 4.4] 시계열 군집분석을 실시한 18개 군집별 시계열 그림

위 [그림 4.4]는 최종적으로 선택된 모형의 18개 군집별 시계열 그림을 나타낸 것이다. Y축은 정규화된 일별 대여량을 나타내고, X축은 일단위의 시간이다. 서로 비슷한 패턴끼리 시계열 군집이 이루어졌으며, 군집별 Y축의 크기에 따라 비슷한 패턴이라도 시계열이 다른 군집으로 나뉘어진 특징이 있다. 아래 [표 4.5]는 군집별 속한 대여소들의 개수이다.

[표 4.5] 군집별 대여소의 개수

| 군집1 | 군집2 | 군집3 | 군집4 | 군집5 | 군집6 | 군집7 | 군집8 | 군집9 |
|---|---|---|---|---|---|---|---|---|
| 4 | 2 | 1 | 32 | 33 | 9 | 7 | 1 | 99 |
| 군집10 | 군집11 | 군집12 | 군집13 | 군집14 | 군집15 | 군집16 | 군집17 | 군집18 |
| 31 | 4 | 64 | 7 | 33 | 28 | 16 | 35 | 34 |

## 4.3 서포트 벡터 회귀모형의 비교

시계열 군집분석이 수요 예측의 정확성을 높이는데 영향을 주는지 비교하기 위해서 두개의 서포트 벡터 회귀모형을 비교하고자 한다. [표 4.6]은 시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형과 그렇지 않은 모형의 비교이다.

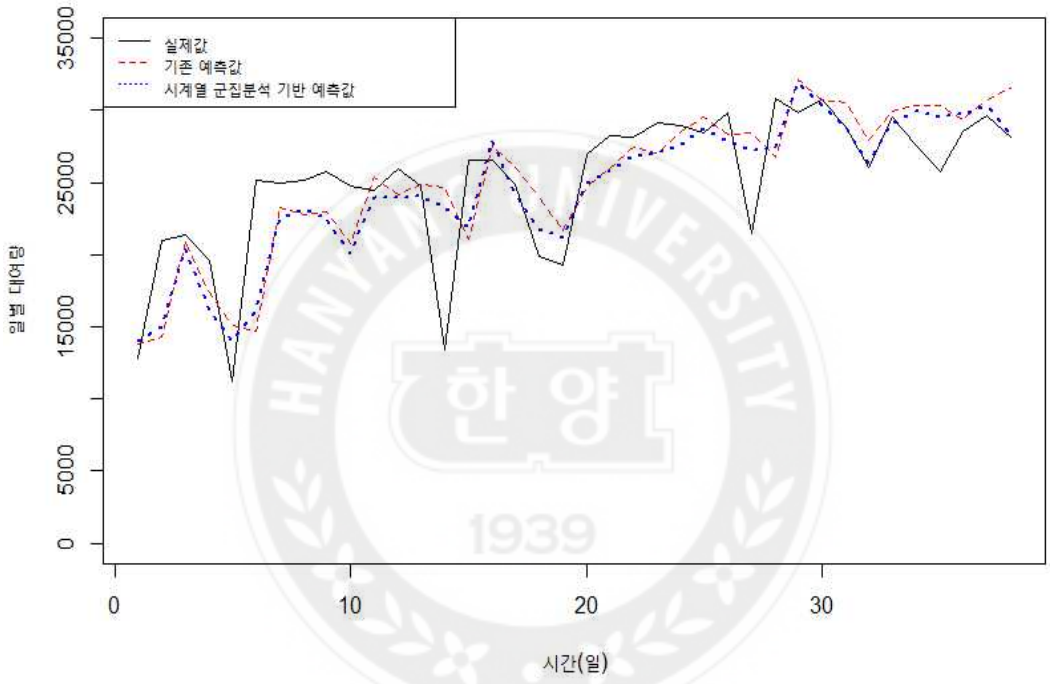
[표 4.6] 시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형과 그렇지 않은 모형의 비교

| 모형 이름 | RMSE | MAE |
|---|---|---|
| 모형 1 | 3728 | 2726 |
| 모형 2 | **3252** | **2399** |

RMSE: 평균 제곱근 오차, MAE: 평균 절대오차

[표 4.6]에서 모형 1은 시계열 군집분석 방법을 사용하지 않은 서포트 벡터 회귀모형이고, 모형 2는 시계열 군집분석 방법을 기반으로 한 서포트 벡터 회

귀모형이다. 평균 제곱근 오차는 3728에서 3252로 12.8% 정도 감소하였고, 평균 절대오차는 2726에서 2399로 12% 정도 감소해 시계열 군집분석 방법을 사용한 모형 2가 더 높은 예측력을 가지고 있다고 판단하였다. 다음 [그림 4.5]는 실제 대여량과 모형1과 모형2의 예측값을 시계열 그림으로 나타낸 것이다.




[그림 4.5] 실제 대여량과 예측값의 비교 시계열 그림

범례의 실제값은 실제 대여량이고, 기존 예측값은 시계열 군집분석을 사용하지 않은 서포트 벡터 회귀모형, 나머지는 시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형의 예측값이다.

# 5. 결론 및 향후 연구과제

본 연구에서는 기존 선행 연구와는 달리 시계열 군집분석과 서포트 벡터 회귀모형의 결합을 이용해 서울시의 공공 자전거 수요 예측을 하였고, 시계열 군집분석을 하지 않은 모형과 비교하였다.

시계열 군집분석을 사용하지 않은 모형은 서울시 전체 대여소의 일별 대여량의 합계를 사용하였고, 시계열 군집분석을 기반으로 한 모형은 서울시 전체 대여소를 모형을 만들기 전 군집분석을 하여 18개의 군집으로 나누었다. 그리고 각 군집별로 최적의 서포트 벡터 회귀모형을 선택하여 예측을 하였고, 그 예측의 합계를 서울시 전체의 자전거 대여량으로 사용해 비교하였다.

분석에 사용된 자료는 2016년 9월 1일부터 2017년 9월 31일까지 총 395일의 자료였다. 과적합 문제를 피하기 위해서 90%의 훈련용 자료의 비율을 이용해 모형을 학습하였고, 10%의 평가용 자료를 이용해 학습된 모형을 평가하였다. 2016년 9월1일부터 2017년 9월1일까지 대략 1년정도의 수요 패턴에 대해 학습하고, 한 달을 예측한 모형이라고 할 수 있다. 모형의 평가는 평균 제곱근 오차와 평균 절대오차를 이용해 평가를 하여 가장 낮은값을 가지는 모형을 선택한 결과, 시계열 군집분석을 한 서포트 벡터 회귀모형이 하지 않은 모형보다 전체적으로 12%정도의 평균 제곱근 오차와 평균 절대 오차가 감소되었다. 분산으로 확인해본다면 시계열 군집분석을 기반으로 한 서포트 벡터 회귀모형의 분산이 시계열 군집분석을 하지 않은 모형보다 실제값의 분산을 10%정도 더 설명하였다.

서울시의 공공 자전거 시스템이 서비스를 시작한지 이제 2년밖에 되지 않아서 많은 양의 데이터를 이용할 수는 없었으나, 향 후 좀 더 기간이 길고, 많

은 자료가 축적된다면 월별, 분기별 등 다양하고 더 정확한 시계열 군집분석과 서포트 벡터 회귀모형을 적용할 수 있을 것이라고 생각한다. 또한 아직 공공 자전거 수요에 대해 예측하는 연구들이 드문 만큼 다양한 분석 방법들을 사용해 더 정확한 수요 예측모형을 구축할 수 있는 여지가 충분한 연구라고 생각된다.

# 참고 문헌


1. 김동준 (2017). “보도자료를 통해 본 주요 교통뉴스”, 한국교통연구원.

2. 김성현 (2015). “서포트 벡터 머신수요예측과 유전알고리즘 기반의 송수펌프 최적운영기법”, 서울시립대학교 환경학과, 석사학위 논문.

3. 김수현, 박승태, 이승철 (2017). “제조 시계열 데이터 처리를 위한 순환 신경망 구조 연구”, *대한 기계학회지*, 2017.2, pp.43-43.

4. 노윤승, 도명식 (2014). “대전시 공공자전거 이용패턴 분석 및 이용수요 예측”, *한국도로학회,* 2014년도 봄 학술대회 논문집, 2014.3, pp.115-18.

5. 문현수, 이영석 (2016). “대전시 공공 자전거(타슈) 공개 데이터 시각화 및 분석”, *한국정보과학회,* 정보과학회 컴퓨팅의 실제 논문지 제22권 제6호, 2016.6, pp. 253-67.

6. 민지원, 문현수, 이영석 (2017). “랜덤 포레스트를 이용한 대전시 공공자전거(‘타슈’) 수요 예측”, *한국정보과학회,* 학술발표논문집, 2017.6, pp.969-71.

7. 신희철, 김동준, 정성엽 (2012). “공공자전거 효과 분석 및 발전 방안”, 한국교통연구원, 기본연구보고서, 2012.10, pp.1-310.

8. 서울시설공단 (http://www.sisul.or.kr).

9. 장다슬 (2016). “서포트 벡터 회귀모형을 기반한 서울지역 가뭄예측”,

한양대학교 대학원 응용통계학과, 석사학위 논문.

10. 최민정 (2016). “딥러닝 알고리즘을 이용한 제주도 관광객 수 예측”, 한양대학교 대학원 응용통계학과, 석사학위 논문.

11. 최희령 (2015). “변동성이 큰 시계열 자료에 대한 신경망 모형과 서포트 벡터 회귀모형 기법의 적용”, 동국대학교 일반대학원, 석사학위 논문.

12. Berndt, D. J. and Clifford, J. (1994). “Using dynamic time warping to find patterns in time series”, In *KDD Workshop*, 10, pp.359-70.

13. Caiado, J., Crato, N., and Pena, D. (2006). “A periodogram-based metric for time series classification”, *Com-putational Statistics & Data Analysis*, 50, pp.2668-84.

14. Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Schölkopf, B., (1998). “Support Vector machines”, IEEE *Intelligent Systems*, Vol.13,No.4, pp.18-28.

15. Hueng-Goo Sohn, Sang-Wook Jung, Sahm Kim. (2016). “A study on electricity demand forecasting based on time series clustering in smart grid”, Department of Applied Statistics, Chung-Ang University.

16. Karatzoglou, A., Meter, D. and Hornik, K. (2006). Support vector machine in R, *Journal of Statistical Softwere*, 15, 1-28.

17. Lu, C. J., T, S. Lee, and C. C. Chiu. (2009). “Financial time series forecatsting using independent component analysis and support vector regression”, *Decision Support Systems*, Vol.47,No.2, pp.115-25.

18. Montero, P. and Vilar, J. A. (2014). “TSclust: An R package for time series clustering”, *Journal of Statistical Software*, 62, pp.1-43.

19. Smola, A. J. and Schölkopf, B. (1998), “A tutorial on support vector regression”, Royal Hollway College, London, U.K, Neuro COLT Tech. Rep.

20. Vapnik, V., Golowich, S. and Smola, A. (1996). “Support vector method for function approximation, regression estimation, and signal processing”, *Advances in Neural Information Processing Systems*, 9, edited by M, Mozer and M, Jordan and T.Petsche, MIT press, Cambridge, MA, 281-87.

# Abstract

## Forecasting public bicycle demand of seoul based on time series clustering

Hanyang University
Department of Applied Statistics
Minhyuk Kim

In this paper we analyzed raw rental data of public bicycle service in Seoul city, using support vector regression after time series clustering. In order to predict the total demand for public bicycles, we compare the models that combine all the rental sites and the models that divide rental sites into clusters. The results of the comparison between the model using the time series clustering and the models that is not used show that the root mean square error and the mean absolute error of the demand forecast support vector regression model based on time series clustering are 12 percent lower. In the time series clustering, the model was selected by comparing the results by distance calculation method and the model was selected by comparing the results for each kernel in the support vector regression. It is expected that the model can be predicted accurately by dividing the rental centers into clusters using time series clustering.

# 연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2017년12월12일

학위명 : 석사

학과 : 응용통계학과

지도교수 : 차경준

성명 : 김민혁 (서명)

**한 양 대 학 교 대 학 원 장 귀 하**

# Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

DECEMBER 12, 2017

| | |
|---|---|
| Degree : | Master |
| Department : | DEPARTMENT OF APPLIED STATISTICS |
| Thesis Supervisor : | Cha Kyung Joon |
| Name : | KIM MINHYUK (Signature) |