한국빅데이터학회지
제8권 제2호, 2023, pp. 97-107 https://doi.org/10.36498/kbigdt.2023.8.2.97

유동인구를 활용한 ConvLSTM AutoEncoder 기반 핫플레이스 탐지

Hot Place Detection Based on ConvLSTM AutoEncoder Using Foot Traffic Data

이주영 · 박헌진†

인하대학교 통계학과

요 약

빅데이터/AI 기반 사회로의 변화에 따른 여러 혜택에서 소상공인은 상대적으로 소외될 가능성이 높다. 이를 지원하기 위해 유동인구를 기반으로 핫플레이스를 정의하여 소상공인의 창업 지역 의사 결정을 지원하고자 한다. 다양한 연구를 통해 해당 지역의 인구 규모가 소상공인의 매출에 중요한 영향을 미친다는 사실이 알려져 있다. 본 연구에서는 인천 유동인구 데이터 중 내륙 지방을 추출하여 연구를 진행하였다. 50m 간격의 격자 형태로 이루어진 데이터로 보간을 통해 일 단위로 이미지화 하였다. LOF와 GAM을 이용하여 공간적 이상치 제거 및 보간을 수행하였고, LOESS를 통해 시간적 이상치를 제거 및 보간하였다. 시간적, 공간적 특성을 모두 고려할 수 있는 ConvLSTM을 예측 모델로 사용하였으며, reconstruction error를 기반으로 이상치 탐지를 수행하는 AutoEncoder 구조를 통해 MAPE가 높은 격자가 밀집해 있는 지역을 핫플레이스로 정의하고자 한다.

■ 중심어 : 유동인구, 핫플레이스, 이상치 탐지, 보간, ConvLSTM AutoEncoder

Abstract

Small business owners are relatively likely to be alienated from various benefits caused by the change to a big data/AI-based society. To support them, we would like to detect a hot place based on the floating population to support small business owners' decision-making in the start-up area. Through various studies, it is known that the population size of the region has an important effect on the sales of small business owners. In this study, inland regions were extracted from the Incheon floating population data from January 2019 to June 2022. the Data is consisted of a grid of 50m intervals, central coordinates and the population for each grid are presented, made image structure through imputation to maintain spatial information. Spatial outliers were removed and imputated using LOF and GAM, and temporal outliers were removed and imputated through LOESS. We used ConvLSTM which can take both temporal and spatial characteristics into account as a predictive model, and used AutoEncoder structure, which performs outliers detection based on reconstruction error to define an area with high MAPE as a hot place.

■ Keyword : Foot Traffic, Hot Place, Anomaly Detection, Imputation, ConvLSTM AutoEncoder

2023년 11월 21일 접수; 2023년 12월 10일 수정본 접수; 2023년 12월 14일 게재 확정.
* 본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행되었습니다.(NRF-2022R1A5A7033499)
† 교신저자 (hjpark@inha.edu)

Ⅰ. 서론

최근 몇 년 동안 도시의 인구 분포의 변화는 빠르게 진행되고 있다. 특히 소상공인은 인구의 움직임과 변화에 민감하게 반응해야 하는데, 이를 효과적으로 파악하고 대응하기 위해서는 “특정 지점을 기준으로 일정 시간동안 이동한 사람의 총 보행량”을 뜻하는 유동인구의 변화에 대한 탐지가 필요하다. 이는 소상공인들에게 중요한 정보를 제공하고, 비즈니스 전략 수립과 시장 파악에 큰 도움을 줄 수 있다. 이임동 외(2010)[1], 이연수 외(2014)[2] 등 많은 연구에서 해당 상권의 인구수가 소상공인의 매출에 유의미한 영향을 미친다는 것이 알려져 있으나 대부분의 경우 인구가 유의미한 영향을 미친다는 것을 발견한 수준에서 머물고 있다. 기계학습 기반 유동인구 추정[3]과 LSTM 기반 유동인구 예측 모형[4]을 제안하는 시도가 있었으나 기계학습 방법은 시간적 상관관계를 파악하지 못하고, LSTM은 공간적 상관관계를 파악하지 못하므로 시간적, 공간적 상관관계가 모두 존재하는 인구데이터 특성상 앞선 방법들에는 한계가 존재한다.

김성아 외(2021)[5]는 핫플레이스를 사람들이 선호하는 음식점과 매장들이 밀집되어 있고 많은 사람들이 방문하게 되는 지역 또는 장소라고 정의하였으며 김태경 외(2018)[6]는 다른 지역과 구별되는 무언가가 존재하며 사람들을 유인하는 요인들로 인해 활기를 띠는 지역이라고 정의하였다. 본 연구에서는 유동인구를 활용한 정량적 평가로 실제값과 예측값 사이의 재구성 오차가 커 예측값에 비해 인구가 빠르게 증가하는 추세를 보이는 격자가 모인 지역을 핫플레이스로 정의하고 이를 탐지하고자 한다.

기존의 핫플레이스 연구 방법은 인구 조사나 통계 자료를 기반으로 하였지만, 이러한 방법은 비용과 시간이 많이 소요되는 문제가 있다. 또한 통계 자료는 대부분 전체 도시 영역에 대한 정보를 제공하기 때문에 세부적인 지역의 핫플레이스를 파악하는 데에는 한계가 있다. 본 연구의 목적은 소상공인을 위해 핫플레이스를 월 단위로 탐지하여 인구 변동에 대한 정확하고 신속한 정보를 제공하는 것이다. 이를 통해 소상공인들은 시장의 동향을 예측하고, 상권 전략을 세우는 데에 도움을 받을 수 있을 것으로 기대된다.

본 연구에서는 격자 형태의 인구 데이터를 활용하여 핫플레이스를 탐지하는 방법을 제안한다. 격자 형태의 인구 데이터는 지역을 작은 격자로 나누어 각 격자 내의 인구수를 기록한 데이터이다. 이를 활용하여 지역별로 인구 밀도를 분석하고, 특정 지역이 핫플레이스로 간주될 수 있는지를 판단하는 모델을 구축하고자 한다. 특히 딥러닝 기법인 Convolutional Long Short Term Memory(ConvLSTM)과 AutoEncoder를 활용하여 인구 데이터를 분석하고, 핫플레이스를 탐지하는 알고리즘을 개발하고자 한다. ConvLSTM은 공간적인 정보와 시간적인 정보를 동시에 모델링할 수 있다는 장점을 가지고 있으며, AutoEncoder는 데이터의 특징을 추출하고 잠재적인 패턴을 학습하는 데에 효과적인 신경망 구조이다.

본 논문은 다음과 같이 구성된다. 2장에서는 전처리 과정에 사용된 통계 모형인 LOF, 일반화 가법모형 및 LOESS와 핫플레이스 탐지에 사용된 딥러닝 모형인 LSTM과 ConvLSTM Auto-Encoder의 이론적 배경에 대해서 설명하고, 3장에서 데이터셋에 대한 소개, 데이터 전처리 과정 및 모델링 과정을 소개한다. 4장은 LSTM과 ConvLSTM 모형 적합 결과를 제시하고 5장에서 결과 정리 및 향후 연구 방향 소개로 마무리 짓는다.

Ⅱ. 이론적 배경

2.1 LOF

Local Outlier Factor(LOF)[7]는 density based anomaly detection 방법 중 하나로, 데이터의 국소적 정보를 이용하여 각 관측치별 이상치 정도를 score로 나타낸다. 다른 객체와의 거리를 계산하여 밀도를 구하며, 현재 포인트의 밀도와 가까운 포인트들의 밀도를 비교하여 현재 포인트의 밀도가 낮을수록 이상치 정도가 높아진다. 전체 데이터의 분포에 영향을 받지 않고 국소적인 정보만을 이용하므로 지역적 패턴을 잘 파악할 수 있다는 장점이 있다.

〈표 1〉 Local Outlier Factor 알고리즘

for $p = 1$ *to* n do
  $k-dist = k-th$ nearest distance at x_p
  $N_k(p) = \{q \in d(p,q) \le k-dist\}$
  for q in $N_k(p)$ do
    $RD_k(p,q) = \max\{k-dist, d(x_i,x_q)\}$
  $lrd_k(p) = \left(\dfrac{\sum_{q \in N_k(p)} RD_k(p,q)}{|N_k(p)|} \right)^{-1}$
  $LOF_k(p) = \dfrac{\frac{1}{lrd_k(p)} \sum_{q \in N_k(p)} lrd_k(q)}{|N_k(p)|}$

$k-dist$는 현재 포인트 x_p와 k번째로 가까운 점과의 거리이며, $N_k(p)$는 x_p와의 거리가 $k-dist$ 이하인 점의 집합이다. 집합 $N_k(p)$에 속하는 원소를 q라 할 때 reachability distance ($RD_k(p,q)$)는 $k-dist$와 x_p, x_q 사이의 거리 중 큰 값을 의미하며, local reachability distance ($lrd_k(p)$)는 $N_k(p)$에 속하는 x_q의 $RD_k(p,q)$의 평균에 역수를 취한 값이다. 마지막으로 $LOF_k(p)$는 $lrd_k(q)$의 평균을 $lrd_k(p)$로 나눈 값으로, 주변 포인트의 밀도와 x_p의 밀도의 비율을 의미한다. $LOF_k(p)$ 값을 기준으로 각 포인트의 이상치 여부를 결정하며 값이 1보다 클수록 뚜렷한 이상치로 판단한다.

2.2 GAM

일반화가법모형(Generalized Additive Model; GAM)[8]은 일반화선형모형(Generalized Linear Model; GLM)을 확장시킨 것으로, GLM에서 독립변수에 대해 적용되었던 선형 관계를 GAM에서는 비모수적 함수를 이용해 비선형적으로 표현할 수 있다. GAM은 가산성은 유지하면서 각 변수의 비선형 함수들을 허용하여 표준선형모델을 확장하는 일반적 체계를 제공한다.

$$Y = f_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \epsilon \sim N(0,\sigma^2) \qquad (1)$$

f_0는 모형의 절편, p는 독립 변수의 개수, $f_j(X_j)$는 unspecified nonparametric function이며, f_j를 추정하기 위해 cubic spline 또는 thin plate spline 등의 비모수적 함수를 사용한다.

이때 $f_j(X_j)$를 추정하기 위해 backfitting algorithm을 사용한다. backfitting algorithm은 모든 $f_j(X_j)$를 초기화하고 $f_j(X_j)$를 제외한 나머지 $f_k(X_k)$를 고정시킨 후 최소제곱합을 만족하는 $f_j(X_j)$를 찾는 방법으로, 아래와 같다.

〈표 2〉 Backfitting 알고리즘

Initialization $f_0 = E(Y), f_1^1 = 0, ..., f_p^1 = 0, m = 0$
for $m := m+1$ do
  for $j = 1$ *to* p do
    $R_j = Y - f_0 - \sum_{k=1}^{j-1} f_k^m(x_j) - \sum_{k=J=1}^{p} f_k^{m-1}(X_j)$
  $RSS = Avg(Y - f_0 - \sum_{j=1}^{p} f_j^m(X_j))^2$
  if $RSS < threshold$ then
    break

2.3 LOESS

선형 최소 제곱 회귀의 단순성과 비선형 회귀의 유연성을 결합한 Local Regression(LOESS)는 관측값 주변의 작은 부분집합에 대해서 회귀 분석을 수행하여 국부적인 추세를 추정한다. 이를 통해 전체 데이터의 전역적인 패턴을 반영하면서도 국부적인 특성을 잘 포착할 수 있다.

LOESS는 주어진 데이터 집합에서 k개의 가까운 데이터를 선택하고, 이를 이용하여 가중치를 계산한다. 일반적으로 LOESS는 가중 평균을 사용하여 회귀식을 추정하는데, 가까운 데이터일수록 더 높은 가중치를 부여한다. 가중 평균을 사용함으로써 이웃들의 관측값이 회귀 분석에 미치는 영향력을 조절할 수 있다.

LOESS는 분포에 대한 가정이 없고 로컬 다항식 회귀를 사용하여 주변 데이터에 따라 관계를 유연하게 조정할 수 있어 비선형적인 관계를 모델링하는 데에 적합하다. LOESS의 알고리즘[9]은 다음과 같다.

〈표 3〉 Local Regression 알고리즘

1. Gather the fraction $s \leftarrow k/n$ of training points whose x_i closest to x_0
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and closest has the hight weight. All but these k nearest neighbors get weight zero.
3. Fit a weighted least square regression of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta_0}$ and $\hat{\beta_1}$ that minimize.

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta_0} + \hat{\beta_1} x_0$

s는 모델 적합에 사용한 x_0 근방의 점의 비율을 나타낸다. s가 커질수록 많은 점을 사용하므로 더욱 smooth해진다. K_{i0}은 x_0와 모델링에 사용될 점 사이의 거리에 따른 가중치를 나타낸다.

2.4 LSTM

LSTM(Long Short Term Memory)[14]은 RNN(Recurrent Neural Network)[15]의 한계 중 하나인 장기 의존성 문제를 극복하기 위해 제안된 모델로 선택적으로 기존 정보를 반영할 수 있도록 설계되었다. 통신사 데이터를 활용한 인구 예측 방법으로 지수 평활법(Exponential Smoothing)[10], ARIMA[11] 등의 고전적인 시계열 예측 방법을 사용할 수 있으나 환경적인 요인(시간 불규칙성, 공간 상관성)에 따라 변동이 큰 데이터를 학습시키기에는 인공지능 모델을 이용한 예측 방법이 더 적합하다. 특히 장기 시계열 예측 문제에서 전통적 시계열 방법론보다 비선형 패턴 또한 파악할 수 있는 LSTM이 좋다고 알려져 있다.[12][13]

LSTM은 입력 게이트(input gate), 망각 게이트(forget gate), 출력 게이트(output gate)와 셀 상태(cell state)로 구성된다. 입력 게이트를 통해 현재 입력 데이터를 얼마나 셀 상태에 추가할지 결정하며 망각 게이트를 통해 이전 셀 상태의 어떤 정보를 잊을지 결정한다. 셀 상태를 통해 현재 타임 스텝의 정보를 저장하고 은닉 상태(hidden state)를 통해 다음 타임 스텝으로 정보를 전달한다. 해당 구조를 통해 장기 의존성을 처리할 수 있는 더 긴 시간 간격에 대한 정보를 유지하고 시계열 데이터에서 중요한 패턴과 관련된 장기 의존성을 캡처하며 기울기 소실 문제를 해결한다. LSTM의 수식은 다음과 같이 표현할 수 있다.

$$i_t = \sigma(W_{xi}x_i + W_{hi}h_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_i + W_{hf}h_{t-1} + b_f)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}x_i + W_{ho}h_{t-1} + W_{co} \circ C_t + b_0)$$
$$h_t = o_t \circ \tanh(C_t)$$

x_t는 현재 시점의 입력 데이터이며 h_{t-1}은 이전 시점의 은닉 상태, C_{t-1}은 이전 시점의 셀 상태이다. $W_{xi}, W_{xf}, W_{xc}, W_{xo}$는 input에 대한 가중치 행렬이며 $W_{hi}, W_{hf}, W_{hc}, W_{ho}$는 은닉 상태에 대한 가중치 행렬, W_{ci}, W_{cf}, W_{co}는 셀 상태에 대한 가중치 행렬, b_i, b_f, b_c, b_o는 각각의 bias이다. $\circ$는 Hadamard product를 나타내며, σ는 시그모이드 활성화 함수를, tanh는 하이퍼볼릭 탄젠트 활성화 함수를 나타낸다.

2.5 ConvLSTM

ConvLSTM[16]은 LSTM과 Convolutional Neural Network(CNN)을 결합한 모델로, 시계열 데이터 및 2D 공간 데이터에 효과적으로 적용할 수 있다. ConvLSTM은 주로 이미지 시퀀스 예측, 동영상 분류, 시계열 데이터 분석 등에 활용된다. ConvLSTM은 LSTM에 컨볼루션 연산을 적용하여 공간 정보를 보존할 수 있도록 개선한 구조이다. 따라서 시간 및 공간적 정보를 함께 다룰 수 있으며, 합성곱 연산을 통해 입력 데이터의 공간 구조를 인식하고 시계열 패턴을 모델링할 수 있다.

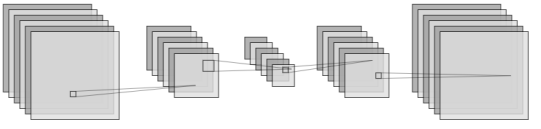
ConvLSTM의 구조는 일반적인 LSTM과 유사하지만, 입력, 은닉 상태, 셀 상태의 차원이 3D 텐서로 구성된다. 일반적으로 ConvLSTM은 2D 입력에 대해 작동하며, 각 시간 단계에서 입력의 각 위치에 대한 은닉 상태와 셀 상태를 유지한다. ConvLSTM의 수식은 다음과 같이 표현할 수 있다.

$$i_t = \sigma(W_{xi} * x_i + W_{hi} * h_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} * x_i + W_{hf} * h_{t-1} + b_f)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo} * x_i + W_{ho} * h_{t-1} + W_{co} \circ C_t + b_0)$$
$$h_t = o_t \circ \tanh(C_t)$$

이때 각 기호는 LSTM과 동일하며 *는 합성곱 연산을 나타낸다.

2.6 AutoEncoder

AutoEncoder는 딥러닝에서의 비지도 학습 방법 중 하나이며 인코더와 디코더로 구성된다. 인코더는 입력 데이터를 저차원의 잠재 공간(latent space)으로 매핑하고, 디코더는 잠재 공간의 표현을 원래 입력으로 복원한다. 이때, 잠재 공간은 데이터의 핵심 특성을 담고 있어 데이터의 차원 축소와 특징 추출에 유용하게 사용된다. AutoEncoder의 가장 큰 특징은 출력을 입력과 같은 형태로 재구성하도록 Loss Function을 최소화하는 것이다. 이러한 특징으로 인해 AutoEncoder는 이상 탐지, 차원 축소, 이미지 복원 등의 다양한 영역에서 활용된다. 본 연구에서는 인코더, 디코더 부분의 각 레이어를 ConvLSTM으로 구성하였다.



〈그림 1〉 ConvLSTM AutoEncoder 구조

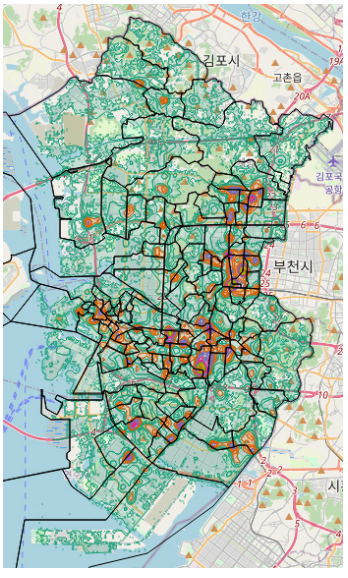
Ⅲ. 연구 내용

3.1 데이터셋

연구에 사용된 데이터셋은 2019년 1월부터 2022년 6월까지의 통신사 생활인구 데이터이며 성연령별 거주인구, 직장인구, 방문인구 등의 정

보가 포함되어 있다. 통신사 생활인구 데이터는 중계기의 신호를 토대로 50m 간격의 격자 내의 인구수를 정제한 데이터이며, 일 단위로 기록되어 있다. 지역의 범위는 인천이며 그 중 섬 지역을 제외하고 일별로 약 80,000~128,000개의 격자(데이터)를 가지는 인천 내륙 지방을 사용하였다. 중계기의 측정 오류, 시그널 정제 과정의 변경 등의 이유로 부분적으로 결측 또는 이상치가 발생하는 격자 등이 포함되므로 통신사 생활인구 데이터에 존재하는 실질적인 격자 개수는 일별로 상이하다.

데이터를 전처리하기 전, 인구의 대략적인 분포를 파악하기 위해 Kernel Density Estimation (KDE)을 이용하여 contour plot으로 시각화한 그림이 아래와 같다. 어두울수록 많은 인구가 몰려 있다는 것을 뜻한다.



〈그림 2〉 KDE contour plot

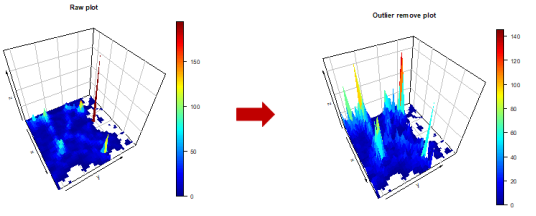
3.2 데이터 전처리

주어진 데이터로 분석을 수행하기에는 공간적 정보를 파악하기 어렵다는 단점이 있다. 따라서 해당 데이터를 2차원 배열의 이미지 형태 데이터로 구조를 변경한 후 분석을 수행하고자 한다. 데이터 전처리 과정은 분석을 수행할 공간 범위 설정, 시간적 공간적 이상치 제거 및 보간 작업을 포함한다. 본 장에서는 핫플레이스 탐지를 위한 학습 데이터셋 생성 과정에서 위의 작업을 수행하는 과정을 설명한다.

이미지 데이터 형태로 변경하기 위해 위도의 범위를 126.58535°에서 126.79313°까지, 경도의 범위를 37.34263°에서 37.63762°까지 설정하였다. 이는 50m 간격으로 368x672 크기의 이미지이며 이를 위해 총 247,296개의 격자가 필요하다. 해당 범위는 바다와 산, 부천 등의 인근 지역을 포함하고 있어 데이터가 존재하는 기간동안 항상 결측인 격자가 존재하며 이에 대해 모두 0을 부여하였다. 해당 기간동안 한 번 이상 데이터가 등장했으나 결측이 발생한 격자는 GAM을 통하여 공간적 보간을 수행하였다. 그 후 각 격자별로 콘서트, 시위 등으로 인해 짧은 기간동안 인구수가 폭발적으로 늘어난 기간을 LOESS를 통해 이상치로 판단하고 제거 후 보간하였다.

3.2.1 spatial anomaly detection

LOF를 이용하여 주변 점들에 비해 비상식적으로 높은 인구수를 기록한 격자를 이상치로 판단하고 해당 값을 제거하였다. LOF는 그 score가 1에 가까울수록 주변 점과의 밀도가 비슷하다는 것을 의미하므로 정상적인 데이터라고 판단하며 1보다 클수록 주변 점의 밀도에 비해 해당 점이 많이 떨어져 있다는 것을 의미하므로 이상치 정도가 크다고 판단한다. LOF score의

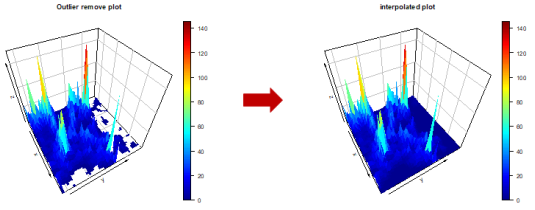


〈그림 3〉 LOF 이상치 제거

이상치 기준에 대한 명확한 통계적 근거는 없으나 일반적인 경우 이상치 탐지 기준으로 1.5에서 2 사이의 값을 사용한다고 알려져 있다. 임계값으로 작은 값을 설정할 경우 정상 데이터까지 이상치로 포함되는 경우가 발생하며 큰 값을 설정할 경우 이상치를 정확하게 탐지해내지 못한다. 본 연구에서는 주변 8개의 격자를 이용하여 LOF를 학습하였으며 각 격자의 LOF score가 2 이상인 점을 이상치로 판단하여 제거하였다.

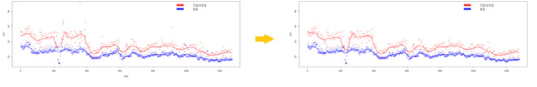
3.2.2 spatial interpolation

기존에 존재하는 결측치와 LOF를 통해 제거된 격자를 x축, y축의 공간적 정보로 인한 비선형적 관계를 잘 잡아낼 수 있는 GAM을 통해 일단위로 적합한 후, 결측 위치를 적합값으로 대체하여 spatial imputation을 수행하였다. 평활 함수로는 thin plate smoothing spline을 사용하였으며 GCV(Genralized Cross Validation)를 통해 smoothing parameter를 설정하였다. 50m 단위의 x,y 좌표를 설명변수로 하여 적합하였다.



〈그림 4〉 GAM spatial imputation

실제 결측이 발생한 격자는 산, 바다 또는 그 인근의 인적이 드문 곳이 대부분이므로 인근 격자의 값이 낮은 경우가 많다. 그에 따라 대부분의 GAM 적합 결과가 0에 가까운 경우가 다수 발생하였다.



〈그림 5〉 LOESS temporal imputaion

3.2.3 temporal anomaly detection

spatial imputation을 완료한 데이터셋을 각 격자 별로 LOESS를 학습하였고, 평일과 주말 간의 차이가 존재하는 것으로 보여 평일과 주말 및 공휴일로 나누어 모델을 적합하였다. 본 연구에서는 LOESS 학습 결과의 잔차를 구하여 잔차의 $\bar{X} \pm 3s$를 기준으로 해당 범위를 초과하는 값을 이상치로 판단하고 제거하였으며 LOESS 결과로 보간하였다.

3.3 모델 적용

앞서 설정한 시간적 공간적 범위 내에서 전처리가 완료된 데이터를 LSTM과 ConvLSTM을 학습하여 RMSE를 비교하고자 한다.

이미지 형태로 학습하므로 값을 0에서 1까지의 범위로 제한하기 위해 min-max scaling을 통해 정규화하였다.

$$x_{scaled} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

두 모델 중 우수한 예측값을 도출하는 모델을 사용하여 재구성 오차를 기반으로 이상치 탐지를 수행하는 AutoEncoder 구조를 통해 예측 결과에 비해 실제값이 높은 지역을 핫플레이스로 정의하고자 한다. 데이터 특성상 절반 이상의 값이 0을 가지며 그에 따라 단순히 RMSE loss function을 사용하게 되면 대부분의 예측 결과가 0으로 편향되어 학습이 제대로 이루어지지 않기 때문에 실제값이 0이 아닌 부분만을 학습하도록 custom RMSE loss function을 정의하였다. 해당 loss function의 수식은 아래와 같다.

$$loss = \sqrt{\frac{1}{\sum_{i=1}^{N} I\{y_i \neq 0\}} \sum_{i=1}^{N} I\{y_i \neq 0\}(y_i - \hat{y_i})^2}$$

Ⅳ. 연구 결과

유동인구 특성상 weekly seasonality가 존재하므로 6주의 같은 요일 데이터를 이용하여 다음 5주의 같은 요일을 예측하도록 데이터셋과 모델을 구성하였다. 2019년 1월 1일부터 2022년 4월 30일까지 368x672 크기의 이미지를 학습 데이터셋으로, 2022년 5월 이미지를 검증 데이터셋으로 하여 RMSE를 계산하였다. 학습 과정 중 모든 epoch는 100으로 고정하였으며 그 외 설정한 매개변수와 예측 결과는 표 4와 같다.

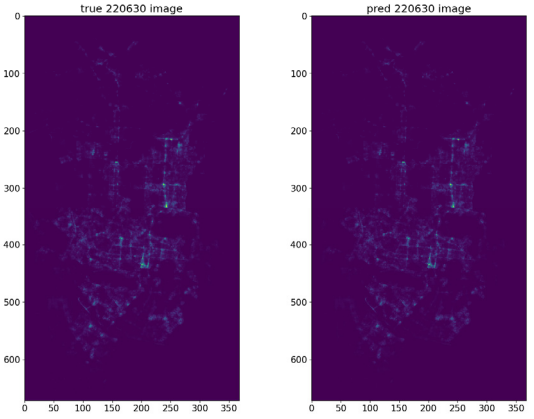
ConvLSTM의 RMSE는 1.3125, LSTM의 RMSE는 3.6426로 ConvLSTM이 LSTM에 비해 보다 나은 예측 결과를 도출하였다.

〈표 4〉 LSTM과 ConvLSTM 결과 비교

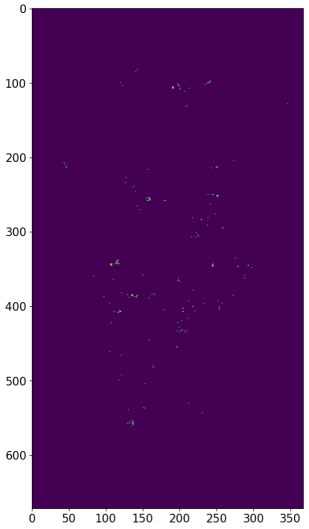
	layers	kernel size	RMSE
LSTM	2	-	3.6427
	4		3.7356
ConvLSTM	2	(3,3)	1.3125
		(5,5)	1.5523
	4	(3,3)	1.5984
		(5,5)	1.6739

예측 결과 가장 좋은 성능을 보인 ConvLSTM AutoEncoder 모델을 이용하여 2022년 6월 1일부터 30일까지의 인구 이미지를 예측한 후 실제값과 함께 나타낸 결과가 그림 6과 같다.

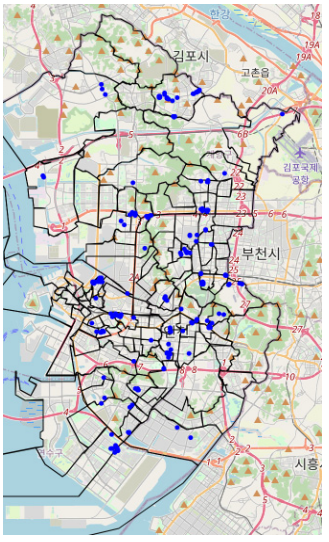
절대적 오차가 아닌 상대적 오차를 통해 격자별 오차의 비율이 큰 격자를 탐지할 수 있는 평가지표인 MAPE(Mean Absolute Percentage Error)를 통해 MAPE가 큰 격자가 밀집해 있는 지역을



〈그림 6〉 22년 6월 30일의 실제 이미지와 예측 이미지



〈그림 7〉 MAPE 기준 상위 200개 격자



〈그림 8〉 MAPE 기준 상위 200개 격자의 실제 위치

탐색하였다. MAPE의 수식은 다음과 같으며 A_t는 실제값, F_t는 예측값을 의미한다.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}|$$

학습 결과 MAPE가 높은 상위 200개 격자에 대한 이미지가 그림 7과 같으며 해당 격자들의 실제 위치를 지도 위에 나타낸 그림이 그림 8과 같다.

MAPE가 높은 격자가 밀집해 있는 지역의 지도 상의 실제 위치를 확인했을 때 잠잠해진 COVID-19로 인해 대면 활동이 활발해진 연세대학교 송도캠퍼스 상권이 인접해있는 송도 캠퍼스타운역 인근, 인구가 증가하고 있는 청라신도시에 접근하기 위해 거쳐가는 가정역 인근, 최근 아파트단지가 완공되며 입주자가 증가하고 있는 검단신도시 신축 아파트단지 등으로 나타났다.

해당 결과는 인천 내륙 지역으로 지역을 한정하고 분석을 수행한 것으로 추후 다른 지역에 대해 적용할 경우 해당 지역의 인구 분포, 지리적 특성 등을 고려하여 추가적인 매개변수 수정 과정이 필요하다.

Ⅴ. 결론

본 연구에서는 2019년 1월부터 2022년 6월까지 일 단위의 통신사 중계기 신호를 정제하여 50m 간격 격자의 중심 좌표와 해당 격자의 인구수를 나타내는 데이터를 사용하여 연구를 진행했으며 인천 내륙 지방으로 그 범위를 설정하였다. 공간 정보를 포함할 수 있는 이미지 형태로 변환하기 위해 범위를 설정하였으며 해당 범위에 산, 바다, 인근 시 등이 포함되고 중계기의 오류로 월별 결측값과 이상치가 존재한다. LOF를 통해 주변 인구수에 비해 값이 매우 큰 이상치를 제거하였으며 온전한 2차원 배열 형태로 만들기 위해 GAM을 통해 공간적 보간을 수행하였다. LOESS를 학습하여 단기간에 인구수가 크게 증가한 기간을 이상치로 판단하여 제거 및 보간하였다. 모델 설계 단계에서는 시간적 공간적 특성을 모두 파악할 수 있는 ConvLSTM을 사용하였으며 AutoEncoder 구조를 통해 reconstruction error가 높은 격자가 밀집한 지역을 예측값에 비해 실제 인구수가 큰 핫플레이스로 정의하였다. 월 단위로 수집되는 데이터를 이용하여 주기적으로 GAM과 LOESS의 매개변수 수정 과정을 거친 후 ConvLSTM AutoEncoder를 재학습하며 모델을 주기적으로 업데이트한다면 보다 정밀한 핫플레이스 탐지가 가능할 것으로 판단된다.

김태경 외(2018)[6] 등 앞선 연구들은 주로 현재 핫플레이스의 위치와 그 특징이 무엇인지에 대해 초점을 맞춘 정성적 연구를 진행하였다. 이에 본 연구는 유동인구에 ConvLSTM AutoEncoder를 학습하여 실제값과 예측값 사이의 재구성 오차를 확인함으로써 인구가 몰리고 있는 지역을 정량적으로 파악하였으며 이러한 딥러닝 기반의 분석 방법을 통해 핫플레이스를 탐지하는 방법을 제안하였다. 이는 추후에 유동인구가 몰리는 지역을 선제적으로 파악하며 소상공인 창업 지역의 의사 결정을 지원하는 하나의 지표로써의 역할을 할 수 있다는 점에서 연구의 의의를 갖는다.

향후 연구에서는 보다 정확한 핫플레이스 탐지를 위해 카드 사용량 데이터와의 융합 방안을 모색할 예정이다. 간단하게는 ConvLSTM의 Convolutional 구조의 channel을 추가하는 방법이 있으며 또는 카드 사용이 발생한 지역만을 데이터로 사용하기 위해 카드 사용 데이터가 발생한 격자를 binary 형태의 2차원 배열로 생성하여 유동인구 데이터에 곱하는 방법을 적용할 수 있다. 또한 모델의 고도화를 위해 유동인구와 카드 사용량 간의 상호작용 관계를 명확하게 분석하여 정밀한 핫플레이스 탐지 방법을 제시할 수 있다.

참 고 문 헌

[1] 이임동, 이찬호, 강상목, “편의점 매출에 영향을 미치는 입지요인에 대한 실증 연구”, 부동산학연구 16(4), pp. 53-77, 2010.

[2] 이연수, 박현신, 유승환, 강준모, “캠퍼스 상권 매출액에 영향을 미치는 입지요인 분석”, 서울도시연구 15(1), pp. 17-34, 2014.

[3] 허찬, 임희석, “기계 학습 기반 유동인구 예측 모델 설계”, 대한 산업공학회 추계학술대회 논문집, pp. 2437-2439, 2018.

[4] 박상윤, 김수현, 허준, “LSTM 기반의 sequence to sequence learning을 이용한 지역 유동 인구수의 예측 모형”, 대한공간정보학회 학술대회, pp. 119-121, 2020.

[5] 김성아, 김흥순, “시멘틱 네트워크 분석을 활용한 ‘핫플레이스’ 의미 분석에 관한 연구”, 국토지리학회지 55(1), pp. 1-13, 2021.

[6] 김태경, 정천용, 정지이, “핫플레이스의 생성조건 및 쇠퇴・이동에 관한 연구”, 경기연구원 기본연구, pp. 1-149, 2018.

[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: identifying density-based local outliers”, in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93-104, 2000.

[8] T. J. Hastie, “Generalized additive models”, in Statistical models in S. Routledge, pp. 249-307, 2017.

[9] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, “The elements of statistical learning: data mining, inference, and prediction”, Springer, 2, 2009.

[10] D. Tikunov and T. Nishimura, “Traffic prediction for mobile network using holt-winter’s exponential smoothing”, in 2007 15th international conference on software, telecommunications and computer networks. IEEE, pp. 1-5, 2007.

[11] H. W. Kim, J. H. Lee, Y. H. Choi, Y. U. Chung, and H. Lee, “Dynamic bandwidth provisioning using ARIMA-based traffic forecasting for Mobile WiMAX”, Comput. Commun. 34(1), pp. 99-106, 2011.

[12] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “A comparison of ARIMA and LSTM in forecasting time series”, in 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp. 1394-1401, 2018.

[13] J. Oliver-Muncharaz, “Comparing classic time series models and the lstm recurrent neural network: An application to S&P 500 stocks”, Finance, Markets and Valuation 6(2), pp. 137-148, 2020.

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory”, Neural Computation 9(8), pp. 1735-1780, 1997.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors”, Nature, 323(6088), pp. 533-536, 1986.

[16] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting”, Advances in neural information processing systems, 28, 2015.

저 자 소 개

이 주 영(Ju-Young Lee)

- 2020년 8월: 인하대학교 통계학과 (이학사)
- 2022년 2월~현재: 인하대학교 통계학과 (이학석사)

<관심분야> 데이터 마이닝, 시계열, 이상치 탐지, 머신러닝

박 헌 진(Heon-Jin Park)

- 1990년 9월~1994년 8월: SAS Institute Inc. Senior Research Statistician
- 1994년~현재: 인하대학교 통계학과 교수, 자연과학대학 학과장

<관심분야> 데이터 마이닝, 시계열, 통계계산