

공간 클러스터링을 활용한 초고속인터넷 해지율 예측 정확도 향상 방안에 대한 연구

박 장 혁*
박 상 언**
김 우 주***

현재 초고속인터넷 시장에서는 통신 3사의 마케팅 비용이 영업이익의 2배가 넘을 정도로 가입자 확보를 위한 마케팅 경쟁이 치열한 상황이다. 그러나 전체적으로 볼 때는 통신사 상호 간에 가입자가 이동하는 구조이기 때문에 시장에 대한 신규가입자 대비 과도한 마케팅비용이 지출된다고 볼 수 있으며, 마케팅 비용을 절약하기 위해서는 신규 가입유치보다 해지 방에 더욱 노력할 필요가 있다. 현재 해지 방은 지역적 특성을 고려하지 못하고 전체 지역에 일반적인 마케팅을 시행하고 있으며, 특히 고객의 해지 예측이 잘 이루어지지 못해 어려움을 겪고 있다. 이러한 상황에서 지역의 특성을 고려하여 해지 예측을 보다 정확하게 함으로써 타겟 마케팅을 할 수 있다면 해지 방에 들어가는 마케팅 비용을 절감할 수 있을 것으로 기대된다.

이를 위해 본 연구에서는 시계열 분석을 통해 초고속인터넷 해지율에 대한 유의미한 예측 가능여부를 확인하고, 해지율 예측 정확도가 높은 최적의 지역 단위를 설정하고자 하였다. 이를 위하여 안양시, 군포시, 의왕시 3개 도시 지역의 47개 행정동의 A통신사 전체 해지 데이터를 수집 및 분석하였다. 해지율 예측 및 검증의 단위를 행정구역 중심의 구와 각 행정동의 공간정보를 기반으로 한 유사 군집으로 나누어 실시해 비교한 결과, 예측 정확도는 군집, 행정구, 전체 순으로 좋게 나타났으며, 전체적으로 해지율 예측이 유의미하다는 결과를 도출하였다. 이는 지역에서 마케팅을 수행할 때 유사 군집 단위로 해지율을 예측하여 추진하는 것이 효과적일 수 있다는 것을 보여준다.

주제어: 공간 클러스터링, 시계열 분석, 초고속인터넷, 해지율 예측, ARIMA

1. 서론

1.1 연구 배경

초고속인터넷 시장에서는 가입자 확대를 위한 통신사들의 마케팅 비용이 매년 증가하고 있다. 과학 기술정보통신부가 제출한 국정감사 자료 <표 1>을 보면 통신사들의 마케팅 비용 지출액이 영업이익의 두

배가 넘는 것으로 나타났다. 주요 원인은 휴대폰 가입자 증대를 위한 경쟁 때문이지만 초고속인터넷(유선인터넷) 가입자 확보를 위한 비용 또한 증가하고 있다. 초고속인터넷의 마케팅비용은 2010년에 발표된 '통신사업자 마케팅비 가이드라인' 발표 이후 감소하였으나 2013년부터 다시 증가추세로 전환되어 2014년에 6,489억, 2017년에는 전년대비 13.5% 증가한 7,776억원을 기록하였다(정보통신정책연구원, 2018).

논문접수일: 2019. 06. 03. 1차 수정본 접수일: 2019. 06. 16. 게재확정일: 2019. 06. 18.

* 연세대학교 공과대학 기술정책 박사과정(janghyuk0212@gmail.com), 제1저자

** 경기대학교 지식정보서비스대학 경영정보전공 교수(supark@kgu.ac.kr), 교신저자

*** 연세대학교 공과대학 산업공학과 교수(wkim@yonsei.ac.kr), 공동저자

무선인터넷 이용이 확대됨에 따라 초고속인터넷의 수요 감소를 전망하는 사람들도 많다. 그러나 과학기술정보통신부에서 발표하는 가입자 현황 자료에 따르면 <표 2>와 같이 휴대폰 가입자뿐만 아니라 초고속인터넷 가입자도 매년 증가추세로 나타났다. 또한, 초고속인터넷 서비스를 기반으로 한 IoT서비스 등 융합형 서비스의 개발이 활발하기 때문에 초고속인터넷은 각종 융합형 서비스를 위한 핵심서비스로 존재가치가 있다고 한다(정보통신정책연구원, 2018).

통신사들은 초고속인터넷의 가입고객증대를 위해 결합상품할인, 요금할인, 위약금 지급, 상품권 지급, 공짜 마케팅 등의 다양한 마케팅 활동을 하고 있다(이상현, 2016). 신규가입 유치 못지않게 중요한 부분이 해지고객 방어이다. 그러나 케이블사업자, 통신사들의 경쟁이 치열하다 보니 신규고객 유치에 쓰이는 노력과 비용대비 해지방어를 위한 마케팅은 적극적으로 하지 못하고 있는 실정이다. 약정이 만료된 고객은 전화 한 통으로 쉽게 해지 할 수 있는 구

조에서 언제 어느 지역의 고객이 해지할지 알 수 없기 때문이다. 포화된 시장에서 한 통신사의 신규고객 유치는 타 통신사의 해지와 직결된다. 실효성 없이 뺏고 빼앗기기를 반복하는 시장 환경이 마케팅 비용의 주된 증가 요인으로 볼 수 있다.

1.2 연구 목적

기존 가입자의 해지방어는 신규가입고객 마케팅 못지않게 중요한 마케팅 요소이다. 그러나 고객의 해지 의사를 예측하지 못하기 때문에 해지 방어 마케팅은 소극적일 수밖에 없다. 이러한 상황에서 해지 여부를 예측할 수 있는 방법이 있다면 보다 효과적인 마케팅 전략을 수행할 수 있을 것으로 판단된다. 개별 고객들의 해지 의사를 직접 예측하지 못한다고 해도, 지역 별로 해지율을 예측할 수 있다면, 해지율이 높아질 것으로 예측되는 지역에 집중적인 가입자 유지 마케팅 실시를 함으로써 가입자를 유지

<표 1> 통신사 영업이익 대비 마케팅비용

(단위 : 억원)

	2015년	2016년	2017년		
	마케팅비용	마케팅비용	마케팅비용(A)	영업이익(B)	(A/B)
SKT	30,547	29,529	31,190	16,977	1.84
KT	26,374	27,142	26,841	9,521	2.82
LGU+	19,640	19,120	21,474	8,437	2.55
합 계	78,319	75,791	79,505	34,935	2.28

*출처 : 금융감독원

<표 2> 연도별 휴대폰, 초고속인터넷 가입자 추이

(단위 : 천명)

구 분		2015년	2016년	2017년	2018년
휴대폰	가입자수	58,935	61,295	63,659	66,356
	증가율	2.9%	4.0%	3.9%	4.2%
초고속인터넷	가입자수	19,818	20,349	20,989	21,286
	증가율	3.2%	2.7%	3.1%	1.4%

*출처 : 과학기술정보통신부 (<http://www.msip.go.kr>)

하고 신규 가입 유치에 투입되는 비용을 절감할 수 있을 것으로 기대된다.

본 연구에는 먼저 시계열 분석을 통해 초고속인터넷 해지율에 대한 유의미한 예측 가능여부를 확인하고, 지역별 마케팅을 보다 수월하게 할 수 있도록 해지율 예측 정확도가 높은 최적의 지역 단위를 설정하고자 한다. 현재 확보할 수 있는 데이터 기준에서 최소의 단위는 행정동이나, 사실상 각 동 별로 다른 마케팅 수단을 사용하는 것은 쉽지 않은 상황이다. 따라서 보다 상위의 행정 단위를 중심으로 해지율 예측과 마케팅 수행을 해야 하는데 구 단위를 선택할 경우, 해지 경향이 서로 유사하지 않은 동들이 구 안에 포함되기 때문에 마케팅의 효율이 떨어질 수 있다. 따라서 본 연구에서는 초고속인터넷의 해지율을 기반으로 특성이 서로 유사한 동들을 군집으로 묶어 해지율을 예측하여 구와 같은 행정 단위 중심의 해지율 예측 결과와 비교해 보고자 한다.

이를 위해 안양시, 의왕시, 군포시 3개 도시 지역의 행정동 단위 47개 동을 대상으로 A통신사의 해지 데이터를 수집하였으며, 행정구역 중심의 구 단위와 유사 군집 지역에 대해 해지율 예측 및 검증 수행한다. 시계열 분석은 해지율 변수만 이용하는 일변량 시계열 모형인 자기회귀누적 이동평균(ARIMA: Auto Regressive Integrated Moving Average)모형을 활용한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 연구와 관련된 기존 연구 내용을 정리하고 3장에서는 연구설계 및 분석방법 이론에 대한 내용을 기술한다. 다음으로 4장에서는 단위 별 시계열 분석을 통해 행정구역 및 군집 단위로 해지 오차율을 도출하고 이에 대한 해석을 서술하고자 한다. 마지막으로 5장에서는 연구요약 및 결론을 기술한다.

II. 관련 연구

2.1 시계열 분석 예측 연구

시계열 분석에 관한 논문은 ARIMA, VAR(Vector Auto Regressive 벡터자기회귀), 지수 평활법(Exponential Smoothing Methods), 이동평균법(Moving average) 등 여러 방법론을 활용한 연구를 쉽게 찾아 볼 수 있다. 그만큼 다양한 분야에서 예측을 통해 미래를 대처하고자 하는 수요가 크다는 것이다. 우선 자산 증대에 대한 기대감 때문인지 부동산에 대한 예측이 많았다. 우경, 이성석(2014)은 1987년부터 2014년까지 전국 지가변동률을 기반으로 단변량 예측 방법인 ARIMA 모형을 활용하여 지가변동률을 예측하였다. 주목할 점은 지가변동률 시계열 데이터에 200만호 주택건설, IMF 외환위기, 미국 서브프라임 모기지 영향을 반영한 개입 ARIMA 모형을 구축하여 세 번의 외부 충격이 전국 지가변동률에 미친 영향을 분석하였다. 김한빈(2017)은 주택시장을 대변하는 핵심 지표인 주택거래량 ARIMA 모형에 인터넷 검색빈도 데이터를 추가한 혼합모형으로 예측하였다. 주택거래량은 특정한 추세 없이 단기 변동성이 큰 정보이기 때문에 국가 정책수립에 있어서 단기예측 및 효과적인 현황 파악이 필요하다. 본 연구에서는 주택거래에 중요한 요소 중 하나인 심리요소를 반영할 수 없는 한계를 극복하고자 인터넷 검색 데이터를 접목했다는 것에 의의를 둘 수 있다. 연구결과 주택거래량 단일변수를 사용했을 때보다 절대평균 오차율을 나타내는 MAPE 기준 약 6% 향상되었음을 설명하였다. 윤주현, 김혜승(2000)은 국토연구원 연구의 일환으로 주택시장 경기동향 및 단기 전망을 하였다. 주택시장은 각종 규제와 금융정책 등으로 지속적으로 재편되고 있고 향후에도 다르지 않을 것이다. 따라서 건설업체, 주택수요자, 주

택정책 당국자 등에게 주택시장 동향 및 전망은 의사결정에 중요한 요소이다. 주택경기를 나타내는 지표에는 주택청약저축가입자수, 주택자금대출, 택지 분양률, 주택 분양률, 부동산중개건수, 취득세 등이 있지만 이 연구에서는 주택가격 단일변수를 기반으로 ARIMA 개입분석모형 및 변환함수모형을 중심으로 단기예측모형을 개발하고 향후 3개월까지의 초 단기 주택시장을 전망하였다. 연구 결과 개입모형에서의 주택 및 전세가격의 단기전망은 서울과 광역시의 3/4분기 전세가격 예측 값을 제외하고 모두 95% 신뢰수준범위로 나타났다. 또한 월별 자료를 통한 예측 값이 분기별 자료의 예측 값보다 나은 결과를 나타냈다. 변환함수모형으로 예측한 결과에서는 2001년의 전국 매매가격은 0.6%정도 증가하고 서울 매매가격은 2.8% 증가하는 것으로 나타났고, 전국 전세가격은 2.9% 증가하고 서울은 5.4% 증가하는 것으로 예측되었다.

철도, 항공, 항만 등 교통, 물류에 대한 시계열 분석 연구도 활발하다. 김관형, 김한수(2011)는 KTX 수요 예측을 위해 개입 ARIMA 모형을 제안하였다. 경부고속철도 2단계 개통과 2008년 금융위기가 경부고속철도 이용량에 미치는 영향을 분석하였다. 금융위기는 통계적으로 유의미한 영향이 없고 경부고속철도 2단계 개통은 평일과 주말에 이용량이 증가한 것으로 나타났다. 윤지성의 3인(2010)은 계절형 시계열 모형을 이용한 국제항공여객 및 화물 수요예측에 대해 연구하였고, 백승한(2006)은 제주-내륙간 국내선 항공여객시장을 대상으로 월별 시계열 총량 자료를 이용하여 계량경제학적 수요 모형을 예측하였다. 이는 항공여객수요 및 운임탄력성의 월별 특성 분석했다는 것에 의의가 있다. 철도의 수요예측은 투자 의사 결정 등에 중요한 정보이기 때문에 예측이 잘못될 경우 과잉투자 및 비용 손실의 문제를 발생시킨다. 김범승(2014)은 중앙선 철도를 중심으로 여객수요를 예측하기 위해 계절 ARIMA 모

형을 제안하여 최적의 모형을 선정하여 예측하였다. 김종길(2011)은 개별항만 발전에 중대한 영향을 주는 컨테이너 물동량을 예측하였는데 ARIMA 모형 및 인과모형중심의 SD모형을 사용하였다. 연구결과 국토부 예측과 차이가 크지 않았으나 예측모형 구축 시 금융위기 등의 사건이 반영되지 못한다는 한계를 보였다.

시계열 분석은 수산물 등 먹거리분야에서도 활용되었다. 남종오, 노승국, 박은영(2012)은 다중회귀 모형, ARIMA 모형, VAR모형을 이용하여 굴 가격 단기 예측을 하였다. 연산 별로 굴 양식은 생산이 불규칙하여 수급불균형을 초래하고 불안정한 가격 추이를 보여왔다. 기존까지 굴의 수급을 체계적으로 분석하여 가격 예측을 하는 연구가 없었지만 본 연구를 통해 연산 별로 예측해봄으로써 향후 굴 위판 가격 이외에 김, 다시마, 미역 등의 예측 참고가 되고 있다.

2.2 공간검색통계량 활용 연구

공간검색 통계량은 1997년 Martin Kulldorff에 의해 제안된 방법이다. 공간검색 통계는 분석대상지역에서 발생한 사건데이터를 활용하여 발생확률이 타 지역에 비해 낮거나 높은 유사지역을 도출하는 방법이다. 발생 데이터를 기반으로 지역별 통계 및 발생 가능 지역을 탐지하는 등 다양한 분야에 사용되고 있다. 박진욱(2016)은 한국의 대기오염수치와 순환기 질환 사망자수를 활용하여 대기오염이 순환기 건강에 미치는 영향을 연구하였다. 공간검색통계는 SatScan 포아송 모델을 활용하였으며 251개 시군구로 분류하여 대기오염과 질환의 발생이 높은 지역을 추정하여 관리 및 예방의 근거를 제시하였다. 황승식(2012)은 공간검색통계량을 이용해 자살률이 높은 군집을 도출하고, 시군구별 자살률을 지도화하여 자살률과 지역의 사회환경적 요인 관련성을 분석

하였다. 또한, 사회환경적 요인과 공간정보에 대한 회귀분석 결과, 기초생활보장가구비율, 물질적 박탈 지수, 정신적 활동제한 인구비율, 초졸 이하 가구 비율, 65세 이상 인구비율이 유의미한 요인으로 분석되었는데 특히 초졸 이하 가구 비율이 높은 지역의 자살률이 상대적으로 높은 것으로 나타났다.

Bambhroliya(2012)는 공간검색통계를 이용하여 유방암 사망률을 연구하였다. 텍사스 주 카운티 단위의 유방암 사망이 혈통 및 인종에 따른 지리적 분포에 미치는 영향을 분석하였다. 연구결과 혈통 및 인종에 따라 유방암 사망률이 높은 다섯 개 군집과 사망위험이 다른 세 개의 군집을 도출하였다. 텍사스 북동부 지역의 군집은 비히스패닉 흑인의 사망이 3.4배로 가장 높은 반면 히스패닉의 사망은 가장 낮았고, 남부지역 군집은 비히스패닉 흑인의 유방암 사망이 가장 낮은 반면 히스패닉의 유방암 사망이 4.77배로 가장 높았다. Chris Green(2003)은 공간검색통계를 통해 캐나다 매니토바주 위니펙스에서 당뇨병 유행의 지리적 변동성, 생태 회귀연구를 하였다. 연구에서 인구의 경제적, 환경적, 생활양식의 다양성과 관련이 있다고 나타났다. 당뇨병 유병률이 높은 지역은 높은 원주민 비율, 낮은 교육 수준, 낮은 가족 수입, 편 부모 가정의 높은 비율, 높은 실업률 등 경제적 지위가 낮고 환경의 질이 좋지 않으며 생활방식이 열악하다는 지표와 밀접한 상관관계를 보였다. 위니펙시의 당뇨병 유병률은 남성보다 여성에서 높았으며 65세 이상 연령층에서 급속하게 증가하는 것으로 나타났다. 상대적 위험도가 높은 클러스터는 위니펙시의 중앙 및 북부 코어에 위치하며, 상대적 위험도가 낮은 클러스터는 도시의 남부 교외에 위치하고 있고 가장 유병 가능성이 높은 클러스터는 도시 중심부로 나타났다. Dahly D.(2011)는 비만 예방을 촉진하기 위해 공유된 환경이 비만 행동에 영향을 미치는지를 이해하기 위해 필리핀 세부(Cebu)지역의 비만에 대한 연구를 하였다. 평균

21.5세의 1,880명의 데이터를 사용하여 공간검색통계를 통해 클러스터링한 결과 메트로폴리탄의 비만률이 높고, 주로 도시지역이 높게 나타났으며 도시 주변과 농촌지역으로 확장되었다.

이 밖에도 멧기열, 고 병원성 조류인플루엔자, 구제역 등 전염성이 높은 질병의 감염 확산, 재발 방지 차원에서 공간검색통계를 이용한 선행 연구들이 있다.

2.3 초고속인터넷 서비스 예측 관련 연구

조상섭, 정동진(2002)은 초고속인터넷 가입자예측력을 비교 평가하였다. 예측기법은 고펜츠(Gompertz) 모형, 단순 로지스틱, 확장 Bass, 시간변동 Bass, 단순성장모형 5가지 예측모형을 사용하였고 예측력은 실측 값 대비 예측 값인 AAD(Average Absolute Difference) 기준으로 예측력을 비교 평가하였다. 연구결과 단순성장모형을 제외한 네 개의 예측모형의 경제적 의미, 통계적 유의성이 유사하게 평가되었다. 예측력은 로지스틱 모형이 가장 높았고 고펜츠 모형이 Bass 모형보다 우월하게 나타났다. 조병선, 고중걸(2002)은 초고속인터넷 가입자의 증가가 둔화될 것이 예상되는 시점에 Bass 모형을 이용해서 Post ADSL의 가입자 수를 예측하였다. 포화점, 확산속도 등을 추정하기 위해서 ADSL, ISDN, CTV에 대한 고펜츠 모형이나 로지스틱을 이용하였다. 추길수(2003) 또한 초고속인터넷과 무선인터넷 서비스의 수요를 Bass모형을 이용하여 수요를 예측하였다. 이러한 예측들은 초고속인터넷 시장이 성숙되는 시점에 기술 경향에 따른 전반적인 시장 예측으로 볼 수 있다.

2.4 기존 연구와의 차이

초고속인터넷 수요 예측에 대한 선행 연구는 2000년대 초반 성장기 시장 예측 수준 정도 이외에는 찾

아 볼 수 없었다. 본 연구는 시계열 분석을 통한 초고속인터넷 해지율 예측이 유의미한 것인지에 대한 연구를 함에 있어서 예측 단위를 행정구 단위, 군집 단위, 전체로 구분하여 예측 정확도가 가장 높은 최적의 예측 단위를 도출 하는 것이 기존 연구와의 차별점이라고 볼 수 있다. 특히, 본 연구는 인접지역을 중심으로 해지율이 유사한 행정동들이 군집으로 묶일 수 있도록 공간정보 기반의 클러스터링을 수행한다는 점에서 학술적 차별점을 갖는다고 할 수 있다. 또한, 초고속인터넷 서비스 마케팅에 관련된 기존 연구가 거의 없는 상태에서 타겟 마케팅이 가능하도록 유의미한 해지율 예측 방안을 제시했다는 것이 본 연구가 갖는 의의이다.

III. 연구 방법

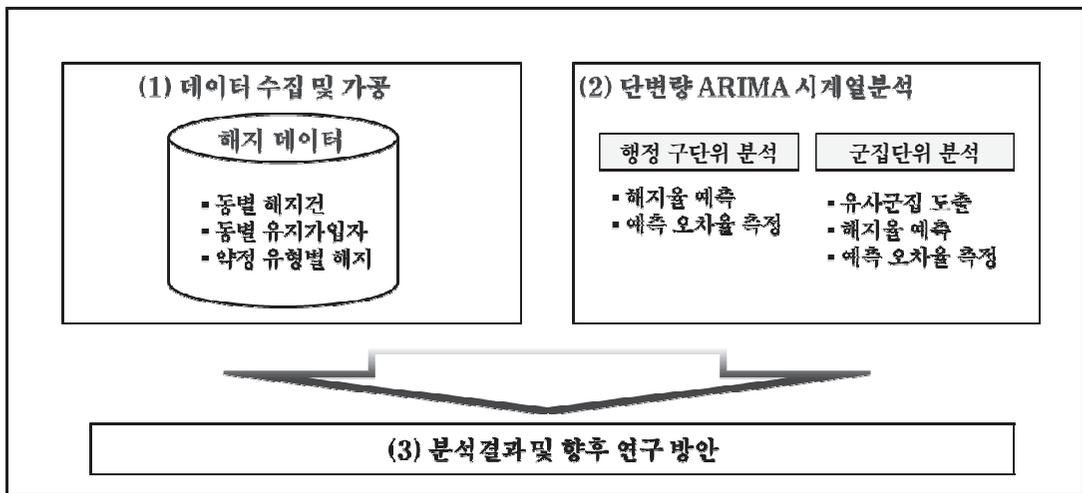
3.1 연구설계

시계열 분석모델을 활용하는 본 연구는 <그림 1>과

같이 크게 3단계로 진행한다. 먼저 (1)데이터를 수집 및 가공하고, (2)해지 데이터를 기반으로 ARIMA 시계열 분석을 통해 해지율 예측 및 오차율을 측정한다. 그 다음 (3)분석결과 및 향후 연구방안을 기술하고자 한다.

초고속인터넷 해지율 예측은 행정구역 단위와 군집 단위의 예측 정확도를 비교하고자 한다. 행정구역 단위는 안양시, 의왕시, 군포시 지역 47개동이 소속된 4개 구 단위과 군집 단위는 공간검색통계를 이용해 해지 유형이 유사한 지역으로 클러스터링된 4개 지역이다. 예측 정확도 측정을 위해 2015년 1월부터 2017년 12월까지 36개월간의 데이터로 예측 모형을 만들고 2018년 1월부터 10개월간을 예측하여 실제 값과 비교하였다. 시계열 분석 방법은 자기회귀누적 이동평균(ARIMA) 모형을 이용하였으며 4개 행정구 단위 및 4개 군집 단위를 각각 분석 및 비교하였다. ARIMA 모형의 식별, 모형의 계수추정, 추정된 모형의 적합성 진단 후 예측하는 순서로 진행하였다.

행정 단위인 구에 비해 보다 유사한 동들로 이루어진 군집에 대한 연구는 이전 연구(박장혁, 2017)



<그림 1> 연구 프로세스

에서 제시된 바 있다. 이 연구에서는 각 동의 초고속 인터넷 결합유형별 해지율과 지리적 위치를 이용하여 인접지역을 중심으로 군집을 도출하고, 각 군집 간에 연령대별 비율, 아파트 평당 매매단가, 자동차 등록비율 등의 특성에서 어떤 차이가 있는지를 분석하였으며, 이 특성의 차이를 기반으로 한 마케팅 방안을 제안하였다. 본 연구에서는 이전 연구에서 제안된 공간정보 기반의 군집분석 방법을 사용하여 초고속인터넷 해지율이 유사한 인접지역에 따른 군집을 도출하고자 한다. 인접지역을 고려한 이유는, 마케팅 수행 시 지역적으로 떨어져 있는 동들보다 군집 단위의 타겟 마케팅이 현실적이기 때문이다.

3.1.1 데이터 수집 및 가공

본 연구에서는 A통신사의 안양시 동안구, 안양시 만안구, 의왕시, 군포시 4개 행정구 단위 47개 동의 해지 데이터를 활용하였다. 기간은 2015년 1월부터 2018년 10월까지 46개월간이며 총 데이터 건수는 63,289건이다. 공간검색통계 도구인 SatScan v9.6의 Poisson모형을 활용하여 군집 단위를 도출하였고 통계 도구인 Eviews10+를 활용하여 ARIMA 시계열 분석 및 예측을 수행하였다.

3.2 자기회귀누적이동평균(ARIMA) 시계열 분석 개요

3.2.1 ARIMA(Auto Regressive Integrated Moving Average) 모형

시간이 흐름에 따라 많은 데이터가 발생한다. 시계열 데이터란 물가지수, 월별판매량, 일일 강수량 등 시간의 흐름에 따라 서로 다른 시점에서 관측된 값을 말한다. 시계열 분석의 목적은 시계열 데이터가 가진 자기종속구조를 파악하여 미래 값을 예측하는 것이다. 시계열 분석 방법에는 지수 평활법, 이동

평균법, 자기회귀누적이동평균(ARIMA)등 다양한 방법이 있다. 본 연구에서는 해지율이라는 변수 하나의 과거 자료를 근거로 미래의 해지율을 예측하는 단변량 ARIMA 시계열 분석 방법을 활용하고자 한다. Box-Jenkins에 의해 체계화된 ARIMA 모형은 단기예측모형으로는 가장 단순하면서 추가 정보 없이 예측이 가능하다는 장점 때문에 쉽게 이용되고 있는 유용한 통계 방법이다(윤주현, 김혜승, 2000).

ARIMA 모형은 AR모형과 MA모형의 혼합모형으로 ARIMA(p,d,q)로 표시하며 다음과 같은 식으로 표기할 수 있다.

- (1) AR(Auto Regressive)모형은 차수가 p인 AR(p)로 표기하며 현시점에서 관측 값 Y_t 를 p개의 과거 관측 값 $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$ 들과 오차항 ϵ_t 의 선형결합형태로 Y_t 자신의 과거 관측 값들에 대하여 회귀시키는 모양이다.

$$Y_t = c + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \dots + \Phi_p Y_{t-p} + \epsilon_t$$

(c는 상수, ϵ_t 는 오차, $\Phi_1 \dots \Phi_p$ 는 AR계수)

- (2) MA(Moving Average)모형은 차수가 q인 MA(q)로 표기하며 현재 Y_t 로부터 일정시점 Y_{t-q} 까지 각각의 관측 값에 같은 가중치를 부여하는 것이라고 할 수 있다.

$$Y_t = c + \epsilon_t - \Theta_1 \epsilon_{t-1} - \Theta_2 \epsilon_{t-2} - \dots - \Theta_q \epsilon_{t-q}$$

(c는 상수, ϵ_t 는 오차, $\Theta_1 \dots \Theta_q$ 는 MA계수)

- (3) ARIMA 모형은 안정적인 시계열 데이터에만 적용된다. 안정적인 시계열 데이터란 분산과 평균에 추세, 계절성 등과 같은 변화 요인이 없는 경우를 의미한다. 따라서 불안정한 경우

로그변환이나 차분을 시행해서 안정적으로 만들어 주어야 한다. ARIMA(p,d,q)의 d는 차분 횟수를 말한다. 차분하여 얻은 시계열 데이터를 W_t 라고 할 때 다음과 같이 표기한다. Δ 는 차분을 의미한다.

$$\begin{aligned}
 W_t &= \Delta d Y_{t-1} \\
 \Delta Y_t &= Y_t - Y_{t-1} \\
 \Delta^2 Y_t &= \Delta Y_t - \Delta Y_{t-1} \\
 W_t &= \Phi_1 W_{t-1} + \Phi_2 W_{t-2} + \dots + \Phi_p W_{t-p} \\
 &+ \epsilon_t - \Theta_1 \epsilon_{t-1} - \Theta_2 \epsilon_{t-2} - \dots \\
 &- \Theta_q \epsilon_{t-q}
 \end{aligned}$$

3.2.2 ARIMA 모형의 구축

Box와 Jenkins는 좋은 ARIMA 모형을 찾기 위해서는 <그림 2>과 같은 절차를 제안하였다. 모형의 구축 절차는 우선 시계열을 안정적으로 만든 후 AR과 MA 차수를 선택하여 모형을 식별하고, 식별된 모형에 포함된 모수들을 추정하며, 추정된 모형의 적합성을 점검하는 모형진단과정을 거친 후 모형이 만족스러운 경우 예측을 하는 순으로 진행한다.

(1) 데이터 정상성

시계열 데이터를 사용할 때 고려해야 할 문제 중 하나는 시계열 데이터의 안정성 여부이다. 따라서 ARIMA 모형식별의 첫 번째 단계는 시계열을 안정적으로 만드는 작업이다. 안정적인 시계열이란 평균

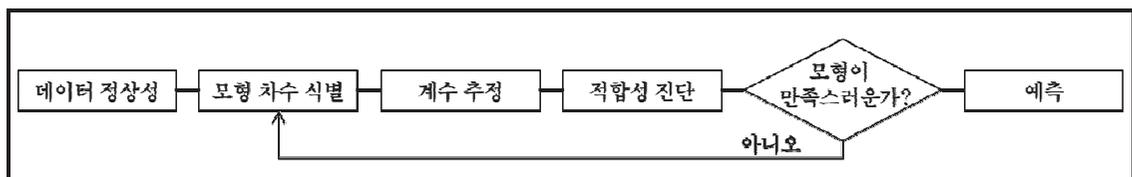
과 분산이 중심으로 크게 벗어나지 않으며 벗어나더라도 그 값이 다시 회복될 것으로 예상할 수 있음을 뜻한다. 불안정한 시계열을 안정적으로 만드는 방법에는 두 가지가 있다. 하나는 차분을 하여 평균을 안정적으로 만드는 방법이고 또 하나는 자연로그를 취하여 분산을 안정적으로 만드는 방법이다. 시계열 데이터의 안정성 검정방법으로는 Augmented Dickey-Fuller 검정과 같은 단위근 검정이 있다.

(2) 모형의 식별

모형 식별은 ARIMA(p,d,q) 모형에서 AR(p)와 MA(q)를 선택하는 단계이다. 이때 표본의 자기상관함수(ACF)와 편 자기상관함수(PACF)의 형태 및 AIC 또는 SBC 정보량을 이용하여 낮은 값으로 AR, MA의 차수를 선택한다. (p: AR의 차수, d: 차분의 수, q: MA의 차수)

(3) 모형의 추정

추정이란 식별된 모형의 계수(parameter)의 유의성을 확인하는 작업이다. 모형의 식별이 AR(p), MA(q) 차수를 결정하는 것이라면 모형의 추정은 AR모형의 p개의 자기회귀계수 $\Phi_1, \Phi_2, \dots, \Phi_p$ 와 MA모형의 q개의 이동평균계수 $\Theta_1, \Theta_2, \dots, \Theta_q$ 의 값의 유의성을 확인한다. Box와 Jenkins는 ARIMA 시계열 모형의 추정 시 최대가능도(ML: Maximum Likelihood)방법을 선호한다. 또한, 안정성과 더불어 ARIMA 모형이 만족해야 할 또 다른 조건인 가역성이 있다. ARIMA 모형은 최근 관측 값일수록



<그림 2> ARIMA 모형의 구축 절차

가중 값이 더 커져야 한다는 전제가 있다. 가역성이 만족되지 않는다는 것은 먼 시점의 관측 값과 가까운 시점의 관측 값이 시점 차이 없이 동등한 가중 값을 갖는다는 의미이다. 가역성 계수가 $|\Theta_1| < 1$ 조건을 만족하면 가역적 모형이다.

(4) 모형의 진단

모형진단이란 추정된 모형이 통계적 성질을 만족하는지의 여부를 판단하는 작업이다. 잔차의 독립성, Ljung-Box Q통계량, 파라미터의 유의성 여부 등으로 진단한다. 우선 잔차의 독립성 여부가 검증되어야 한다. 올바른 모형이라면 독립적으로 계열상관이 없어야 하기 때문에 일반적으로 잔차를 산출하고 먼저 잔차가 독립인지 아닌지에 대한 통계적 검정을 행하게 된다(윤주현, 2000). 잔차 자기상관을 검증하는 방법에는 Box-Pierce Q통계량과 Ljung-Box Q통계량 등이 있다.

3.3 유사 군집 도출

본 연구에서는 공간검색통계 기반의 군집분석 방법을 사용하여 초고속인터넷 해지율이 유사한 인접 지역에 따른 군집을 도출하고자 한다. 군집은 각 동의 위치 값과 해지율 데이터를 기반으로 하여 공간 검색통계 도구인 SatScan의 포아송 모형을 활용하여 추출하였다. 포아송 분포는 불규칙하게 발생하는 확률을 표현하는 이산확률분포이며 거리, 공간, 시간 등에서 무작위로 발생하는 사건의 확률을 구하는데도 사용한다. 확률에 따라 사건이 밀집된 지역과 분산된 지역이 구분될 수 있는데, 포아송 분포는 이를 파악하는 기준이 될 수 있다(유철상, 2011).

SatScan의 공간 클러스터링은 각 지역의 좌표 값을 중심으로 분석대상지역의 반지름이 50%가 될 때까지 확대하며 원 내부의 우도 비율 값과 원 외부의 우도 비율 값의 차이를 기준으로 군집을 도출한

다(한준희, 2016). 우도 비율 값의 차이가 크다는 것은 우도 비가 높다는 것이고 해당 군집 내의 발생 빈도가 그 이외 지역보다 높기 때문에 하나의 군집으로 볼 수 있다는 것을 의미한다(이경주, 2015). 통계적 클러스터는 SatScan의 우도 비에 따라 표시되는데 포아송 모형의 우도 비는 다음과 같다.

$$Y = \left(\frac{C}{E[C]} \right)^c \left(\frac{C-c}{C-E[c]} \right)^{c-c} I_0$$

C : 전체 값, c : 측정 값, E(c) : 측정 값의 기대 값,
I() : 지시함수

IV. 단변량 시계열 분석을 이용한 해지율 예측

초고속인터넷 해지율 예측은 행정구역 구 단위와 군집 단위의 예측을 비교하고자 한다. 행정구역 구 단위는 안양시 동안구, 안양시 만안구, 군포시, 의왕시 이상 4개 구이고 군집 단위는 해지율이 유사한 지역으로 클러스터링된 4개의 군집이다.

4.1 구 단위 해지율 예측

구 단위는 <표 3>와 같이 47개 행정동이 속한 4개 구 단위로 분석하였다. 본 논문에서는 안양시 동안구를 예로 들어 ARIMA 모형의 식별, 추정, 검증, 예측의 과정으로 설명하였다.

4.1.1 시계열 데이터의 정상화

ARIMA 모형은 시계열 데이터가 가지고 있는 추세, 계절, 주기 등 불규칙한 요인들이 제거된 안정적인 데이터여야 한다. 따라서, 불안정한 시계열 데이

〈표 3〉 구 단위 지역 정보

구 분	인구	면적	행정동수
안양시 동안구	332,787	22 Km	17 개
안양시 만안구	254,977	37 km	14 개
군포시	271,179	36 Km	11 개
의왕시	155,767	54 Km	05 개

*출처 : 통계청

터를 안정적으로 만드는 것에는 두 가지 방법이 있다. 하나는 평균이 일정하지 않을 경우 차분을 행함으로써 평균을 안정적으로 만드는 방법이며, 다른 하나는 분산이 일정하지 않을 경우 자연로그를 취해 분산을 안정적으로 만드는 방법이다. 동안구 시계열 데이터는 변동 폭이 불안정하여 로그를 취해 안정성을 확보했다.

시계열 데이터의 정상성을 확인하는 방법에는 단위근 검정이 있다. 파라미터가 1에 가까운 값을 가지면 단위근이 존재한다는 것이며 이 경우 해당 ARIMA 모형을 통한 예측은 의미가 없다(우경, 2014). 로그변환을 취한 동안구 시계열 데이터를 Augmented Dickey-Fuller 검정을 통해 단위근 유무를 확인해본 결과 p-값이 0.19%로 안정적인 시계열인 것으로 나타났다. 즉, ‘단위근이 존재한다’ 또는 ‘주어진 시계열의 평균이 비정상적이다’라는 귀무가설을 기각한다.

4.1.2 모형의 식별

ARIMA 모형은 $ARIMA(p,d,q)$ 로 표시하는데 모형식별은 $AR(p)$ 과 $MA(q)$ 차수를 결정하는 것이다. p 는 AR 의 차수, d 는 차분한 횟수, q 는 MA 의 차수를 뜻한다. 동안구 시계열 데이터는 차분을 하지 않았으므로 $d=0$ 인 $ARIMA(p,0,q)$ 모형이다.

차수 p 와 q 는 자기상관함수(ACF)와 부분자기상관함수(PACF)형태를 보고 결정할 수 있다(윤주현, 2000). 하지만 실제 자기상관함수와 부분자기상관함수를 그려보면 판별기준이 뚜렷하게 나타나지 않고 판단하기 난해한 경우가 많아서 적절한 p 와 q 를 찾기 어렵다. 그래서 일반적으로 AIC(Akaike Information Criterion), SBC(Schwartz-Bayesian Criterion), HQ(Hannan-Quinn information criterion)등의 정보기준 통계량을 이용하여 모형을 선택한다(김관형, 2011). 차수선택은 〈표 4〉에 보여지는 것과 같이 각 조합의 통계량을 확인하여 AIC, SBC, HQ가 가장 낮은 $ARIMA(0,0,1)$ 조합을 선

〈표 4〉 ARIMA모형별 정보기준 통계량

Model	AIC	SBC	HQ
(0,0,1)	-1.43700	-1.30504	-1.39094
(1,0,0)	-1.43238	-1.30042	-1.38632
(2,0,0)	-1.38657	-1.21063	-1.32516
(0,0,2)	-1.38616	-1.21021	-1.32475
(1,0,1)	-1.38453	-1.20858	-1.32312

택하였다.

4.1.3 계수 추정 결과

안정성과 더불어 ARIMA 모형이 만족해야 할 또 다른 조건인 가역성이 있다. 가역성이 만족되지 않는다는 것은 먼 시점의 관측 값과 가까운 시점의 관측 값이 시점 차이 없이 동등한 가중 값을 갖는다는 의미이다. 가역성 계수가 $|\Theta_1| < 1$ 조건을 만족하면 가역적 모형이다.

가역성 여부 검정결과 가역성 값이 0.3384로 $|\Theta_1| < 1$ 조건을 만족하는 가역적 모형으로 확인되었다. 또한 ARIMA(0,0,1)모형의 추정된 계수의 통계적 유의성 검정결과인 <표 5>에 있는 t통계량을 보면 상수항은 0.1% 수준에서 유의미하고 MA(1)은 10% 수준에서 유의미한 것으로 나타났다.

4.1.4 모형 진단 결과

모형 진단이란 추정된 모형이 통계적 성질을 만족하는지의 여부를 판단하는 작업이다. 잔차의 독립성, Ljung-Box Q통계량, 파라미터의 유의성 여부 등으로 진단한다.

우선 잔차의 독립성 여부가 검증되어야 한다. 올바른 모형이라면 독립적으로 계열상관이 없어야 하기 때문에 일반적으로 잔차를 산출하고 먼저 잔차가 독립인지 아닌지에 대한 통계적 검정을 행하게 된다(윤주현, 2000). 잔차 검정을 시행해본 결과 추정

된 ARIMA(0,0,1)모형 잔차의 자기상관함수 및 편자기상관함수가 신뢰선 수준 안에 있기 때문에 적합한 모형으로 볼 수 있다. 또한, 각 차수별 Ljung-Box Q통계량에서도 잔차의 독립성이 검증되었다.

4.1.5 모형 예측 결과

추정된 ARIMA 모형을 통해 도출한 2018년1월에서 2018년10월까지 10개월간의 예측 값과 실측 값을 비교하였다. 예측 적합도는 MAPE(Mean Absolute Percentage Error)를 활용하였다. MAPE는 절대평균오차율을 말하는데 모형의 예측 정확도를 판단하는 기준 중 하나이고 작은 값이 예측력이 높음을 나타낸다. MAPE의 적합도 계수는 10% 미만일 때 매우 정확한 예측, 20% 미만일 때 비교적 정확한 예측, 50% 미만일 때 비교적 합리적인 예측으로 보며 50%이상일 경우 예측이 부정확하다고 해석할 수 있다(윤한영, 2019).

$$MAPE = \frac{1}{n} \sum \left| \frac{X_t - F_t}{X_t} \right|$$

X : 관측 값, F : 예측 값, n : 관측기간(개월수)

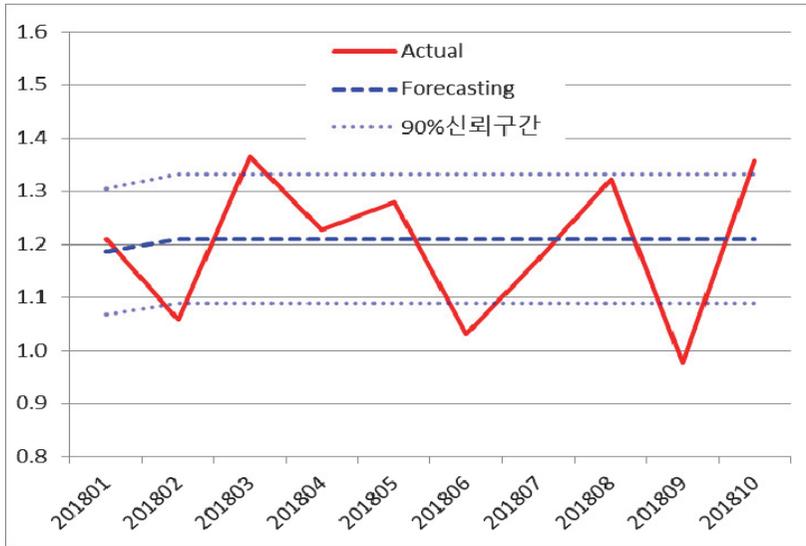
ARIMA(0,0,1) 모형을 통한 동안구 예측은 <그림 3>에서 보이는 바와 같이 평균으로 수렴하는 그래프를 나타내고 있다.

예측 값을 동안구에 속한 17개 행정동 별 실측 값과 비교한 결과 동안구의 평균 예측 오차율 MAPE

<표 5> ARIMA(0,0,1)모형의 계수

Variable	Coefficient	Std. Error	t-Statistic
C	0.190924	0.028252	6.757901***
MA(1)	0.338495	0.189998	1.781574 ⁺
R-squared : 0.115550		Adj. R-squared : 0.061947	

⁺:p<0.1, **:p<0.05, ***:p<0.01, ****:p<0.001



〈그림 3〉 동안구 해저울 예측결과 (2018.1~2018.10)

〈표 6〉 행정동 별 예측 오차율

만안구		군포시		동안구		의왕시	
석수3동	10.6%	군포2동	11.2%	호계2동	9.3%	내손1동	25.2%
박달2동	12.3%	산본1동	13.0%	부림동	13.8%	부곡동	25.6%
석수2동	12.4%	산본2동	14.7%	호계3동	14.0%	내손2동	25.8%
석수1동	14.3%	금정동	16.6%	관양1동	15.9%	오전동	30.2%
안양6동	14.4%	광정동	30.8%	관양2동	15.9%	고천동	48.9%
안양9동	15.2%	군포1동	31.2%	비산1동	16.6%	청계동	51.7%
안양3동	15.4%	궁내동	36.0%	비산2동	17.2%		
안양1동	15.6%	오금동	44.9%	평촌동	17.4%		
박달1동	15.8%	채궁동	67.3%	비산3동	19.7%		
안양5동	16.0%	수리동	97.3%	호계1동	24.3%		
안양8동	16.1%			갈산동	40.1%		
안양2동	17.4%			범계동	40.1%		
안양7동	20.7%			부흥동	41.1%		
안양4동	40.2%			달안동	51.9%		
				평안동	86.7%		
				귀인동	102.2%		
				신촌동	143.9%		
가중평균	16.0%	가중평균	19.1%	가중평균	21.4%	가중평균	30.2%

는 21.4%이고 행정동 별 예측 오차율은 17개동 중 13개동이 40% 미만으로 비교적 합리적인 예측으로 나타났으며 그 중 9개동은 20% 미만으로 비교적

정확한 예측이 된 것으로 나타났다.

위와 동일한 방법으로 다른 구 단위 지역의 예측 오차율을 시행한 결과는 〈표 6〉과 같다. 만안구가

예측 오차율 16.0%로 예측 정확도가 가장 높았고 군포시 19.1%, 동안구 21.4%, 의왕시 30.2% 순으로 나타났다.

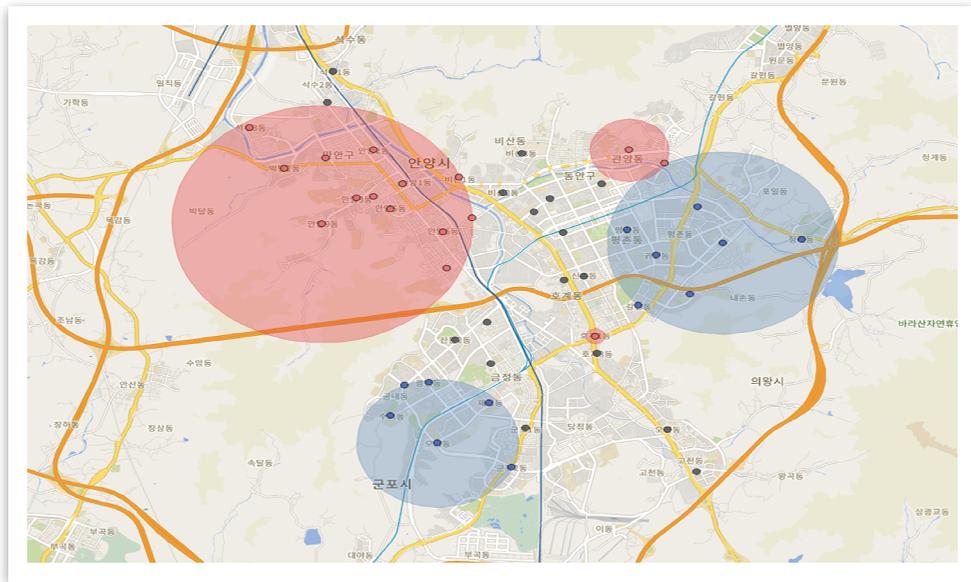
4.2 군집 단위 해지율 예측

4.2.1 군집 도출

47개 행정동 별 36개월간(2015년~2017년)의 해지율 데이터를 기반으로 공간검색통계를 통해 <그

림 4>과 같이 5개의 군집을 도출하였다. 공간검색통계는 SatScan의 Poisson 모형을 활용하였으며 본 연구가 해지율 단일변수만을 활용하는 ARIMA 분석이기 때문에 군집분석도 해지율 단일변수만을 사용하였다.

도출된 총 5개 군집의 정보는 <표 7>과 같다. 해지율 예측은 5개 군집 중 1개 동으로 형성된 군집을 제외한 4개 군집에 대해 실시하였다. 본 논문에서는 4개 군집 중 행정동이 가장 많은 군집 ④를 예로 들어 ARIMA 모형의 식별, 추정, 검정, 예측의 과정



<그림 4> 군집 형성도

<표 7> 군집 별 클러스터링 정보

군집	동수	행정동명	p-value
①	2개	광양1동, 광양2동	0.0000000
	1개	호계1동	0.0000048
②	6개	오금동, 수리동, 재궁동, 광정동, 군포2동, 궁내동	0.0000000
③	7개	내손2동, 평촌동, 귀인동, 내손1동, 청계동, 평안동, 갈산동	0.0000027
④	13개	안양9동, 안양3동, 안양4동, 안양5동, 박달2동, 박달1동, 안양1동, 안양2동, 안양6동, 안양8동, 석수3동, 비산1동, 안양7동	0.0017000

으로 설명하였다.

4.2.2 시계열 데이터의 정상화

ARIMA 모형은 시계열 데이터가 가지고 있는 추세, 계절, 주기 등 불규칙한 요인들이 제거된 안정적인 데이터여야 한다. 군집 ④의 시계열 데이터는 평균이 일정하지 않아 차분을 시행함으로써 안정성을 확보했다.

Augmented Dickey-Fuller 검정을 실시해본 결과 p-값이 0.00% 수준으로 단위근이 존재하지 않는 안정적인 시계열로 나타났다. 즉, '단위근이 존재한다' 또는 '주어진 시계열의 평균이 비정상적이다'라는 귀무가설을 기각한다.

4.2.3 모형의 식별

ARIMA 모형은 $ARIMA(p,d,q)$ 로 표시하는데 모형추정은 $AR(p)$ 과 $MA(q)$ 차수를 결정하는 것이다. p 는 AR 의 차수, d 는 차분한 횟수, q 는 MA 의 차수를 뜻한다. 군집 ④는 1차 차분을 했으므로 $d=1$ 인 $ARIMA(p,1,q)$ 모형이다. 차수 p 와 q 는 일반적으로 AIC, SBC, HQ 등의 정보기준 통계량을 이용하여 모형을 선택한다. 차수선택은 <표 8>에 보여지는 것과 같이 각 조합의 통계량을 확인하여 AIC 및 HQ가 가장 낮은 $ARIMA(2,1,3)$ 조합을 선택하였다.

4.2.4 계수 추정 결과

ARIMA 모형이 만족해야 할 또 다른 조건인 가역성 여부 검정 결과, $ARIMA(2,1,3)$ 의 가역성 값은 전부 $|\Theta_1| < 1$ 조건을 만족하는 가역적 모형으로 확인되었다. 또한 모형의 추정된 계수의 통계적 유의성 검정결과는 <표 9>와 같이 상수항의 t통계량은 5% 수준에서 유의미하고 $AR(2)$ 의 t통계량은 0.1% 수준에서 유의미한 것으로 나타났다.

4.2.5 모형 진단 결과

추정된 모형이 통계적 성질을 만족하는지의 여부를 판단하는 작업은 일반적으로 잔차를 산출하고 먼저 잔차가 독립인지 아닌지에 대한 통계적 검정을 행하게 된다(윤주현, 2000). 잔차 검정을 시행해본 결과 추정된 $ARIMA(2,1,3)$ 모형 잔차의 자기상관함수 및 편 자기상관함수가 신뢰선 수준 안에 있기 때문에 적합한 모형으로 볼 수 있다. 또한, 각 차수별 Ljung-Box Q통계량에서도 잔차의 독립성이 검정되었다.

4.2.6 모형 예측 결과

추정된 $ARIMA(2,1,3)$ 모형을 통해 추출한 예측 값과 실측 값 비교그래프는 <그림 5>와 같다. 군집 ④의 예측 오차율 MAPE는 17.8%이고 예측 값을

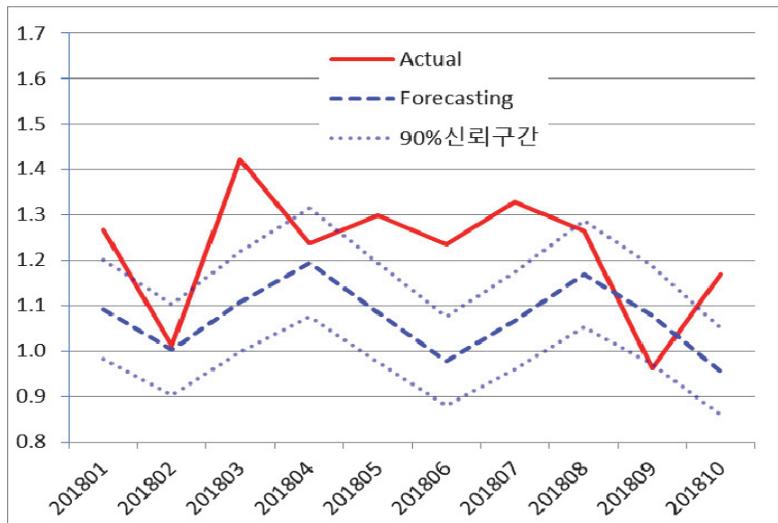
<표 8> ARIMA모형별 정보기준 통계량

Model	AIC	SBC	HQ
(2,1,3)	-1.0131	-0.70203	-0.90572
(3,1,3)	-0.97224	-0.61673	-0.84952
(2,1,4)	-0.96496	-0.60946	-0.84224
(0,1,1)	-0.9301	-0.79678	-0.88408
(2,1,1)	-0.92201	-0.69981	-0.84531

〈표 9〉 ARIMA(2,1,3)모형의 계수

Variable	Coefficient	Std. Error	t-Statistic
C	-0.006004	0.002926	-2.051697*
AR(1)	0.08817	0.059260	1.487842
AR(2)	-0.998369	0.037961	-26.30014***
MA(1)	-1.046479	663.9493	-0.001576
MA(2)	1.046482	992.5638	0.001054
MA(3)	-0.999966	1569.8200	-0.000637
R-squared : 0.64338		Adj. R-squared : 0.566961	

+:p<0.1, *:p<0.05, **:p<0.01, ***:p<0.001



〈그림 5〉 Cluster④ 해지율 예측결과 (2018.1~2018.10)

군집 ④에 속한 13개 행정동 별 실측 값과 비교한 결과 13개동 모두 예측 오차율 40%미만으로 비교적 합리적인 예측이 되었으며 그 중 10개동은 20% 이하인 비교적 정확한 예측으로 나타났다.

위와 동일한 방법으로 다른 군집의 예측 오차율을 시행한 결과는 〈표 10〉과 같다. 동안구 관양1,2동이 속한 군집 ①의 예측 오차율이 11.8%로 정확도가 가장 높았고 만안구지역에 형성된 군집 ④는 17.8%, 군포지역에 형성된 군집 ②는 22.0%, 의왕시와 동안구 지역에 형성된 군집 ③은 31.1%

로 나타났다.

4.3 전체 해지율 예측

전체 해지율 예측에서는 위와 동일한 방법으로 전체 단위 즉 모든 동을 합하여 하나의 단위로 만들어 해지율 예측을 수행하였다. 즉 전체 단위는 안양시 동안구, 안양시 만안구, 군포시, 의왕시 4개 구 단위의 47개 행정동을 모두 포함한 전체 지역이다. 전체 단위의 시계열 데이터는 분산이 불안정하여 로그를

〈표 10〉 군집별 예측 오차율

군집 ①		군집 ④		군집 ②		군집 ③	
관양2동	11.4%	석수3동	12.9%	군포2동	13.6%	갈산동	19.8%
관양1동	12.0%	안양3동	13.9%	광정동	22.2%	청계동	25.3%
		비산1동	14.2%	궁내동	31.5%	평촌동	25.6%
		안양5동	15.1%	오금동	38.1%	귀인동	33.0%
		박달2동	16.0%	재궁동	63.9%	내손2동	35.3%
		안양6동	16.2%	수리동	91.1%	내손1동	35.7%
		안양9동	16.9%			평안동	39.2%
		안양2동	17.0%				
		안양8동	18.2%				
		안양1동	19.1%				
		박달1동	21.4%				
		안양7동	24.6%				
		안양4동	37.4%				
가중평균	11.8%	가중평균	17.8%	가중평균	22.0%	가중평균	31.1%

취해 안정성을 확보했으며 ARIMA 모형은 AIC, SBC, HQ 통계량을 이용하여 ARIMA(0,0,4) 조합을 선택했다. ARIMA(0,0,4) 모형을 통한 전체 해지율 예측 값을 47개 행정동별 실측 값과 비교한 결과, 전체에 대한 예측 오차율은 22.9%로 군집 단위, 구 단위보다 예측 정확도가 낮게 나타났다.

4.4 분석결과 토의 및 마케팅 방안 도출

본 연구에서는 구 단위, 군집 단위, 전체 총 3개의 지역 단위로 구분하여 시계열 분석을 통한 해지율을

예측했다. 〈표 11〉을 보면 알 수 있듯이 군집 단위의 예측오차율 평균은 20.7%이고 구 단위 예측오차율 평균은 21.7%로 산출되어 군집 단위의 예측 정확도가 더 높게 나타났다. 연간 초고속인터넷 해지건수는 약 3백만건(초고속인터넷 가입자 2천만 X 연평균 해지율 15%)에 달한다. 이중 예측 오차율 1% 차이는 3만건에 해당하고 마케팅비용 건당 40만원으로 산정하여 금액으로 환산하면 연간 120억에 해당되는 결코 작은 금액이 아니다. 이번 분석결과를 통계 및 마케팅 시행 시 현재와 같이 행정구역 단위로만 추진하는 것 보다 유사 군집 단위로 추진

〈표 11〉 지역 단위 별 예측 오차율 비교

		동안구	만안구	군포시	의왕시	계
구 단위	관측 동수	17	14	10	6	47
	예측 오차율	21.4%	16.0%	19.1%	30.2%	21.7%
		군집1	군집2	군집3	군집4	계
군집 단위	관측 동수	2	6	7	13	28
	예측 오차율	11.8%	17.8%	22.0%	31.1%	20.7%

〈표 12〉 군집 별 지역 분포 현황 (단위: 동 개수)

	동안구	만안구	군포시	의왕시	계
군집 ①	2				2
군집 ②			6		6
군집 ③	4			3	7
군집 ④	1	12			13
계	7	12	6	3	28

〈표 13〉 예측 순위 비교 (단위: 동 개수)

	동안구	만안구	군포시	의왕시	계
구 단위 예측 순위 동수	1	8	1	2	12
군집단위 예측 순위 동수	6	4	5	1	16
계	7	12	6	3	28

하는 것이 효과적일 수 있다는 것을 보여준다.

〈표 12〉를 보면 군집으로 형성된 행정동은 총 28개동이며 군집 ①은 동안구 2개동, 군집 ②는 군포시 6개동, 군집 ③은 동안구 4개동과 의왕시 3개동, 군집 ④는 동안구 1개동과 만안구 12개동으로 구성되어 있다. 이렇게 군집 별로 해당되는 행정구가 2개 이상인 경우가 있기 때문에 구 별 예측률과 군집 별 예측률의 1:1 비교는 무의미하다.

다만 각 동 별 예측률 비교는 가능하다. 〈표 13〉은 군집으로 형성된 28개 행정동을 대상으로 구 단위 예측 값을 적용했을 경우와 군집 단위 예측 값을 적용했을 때를 비교한 결과이다. 총 28개 동 중 16개 동이 군집 단위를 적용했을 때 예측 정확도가 높았고, 12개 동은 구 단위 예측 값을 적용했을 때 높게 나타났다. 두 예측 단위의 동별/월별 예측 오차율에 대한 t검정에서도 p-값이 0.01% 수준에서 유의미하여 군집단위의 예측 정확도가 높다고 볼 수 있다.

군집의 특성을 살펴보면 군집 ①과 군집 ④는 해지율이 높은 행정동의 집단이고 군집 ②와 군집 ③은 해지율이 낮은 행정동으로 구성되었다. 해지율이

높은 집단과 낮은 집단의 마케팅은 다르게 적용될 필요가 있다. 높은 집단은 약정 만료시 타사로 쉽게 이동한다는 의미이고 낮은 집단은 A통신사에 충성도가 높다고 볼 수 있다. 또한 해지율이 높은 집단이 경제적으로 부유하지 않을 수도 있다. 이렇게 고객의 구성 및 특성이 다양하기 때문에 관행적으로 시행하던 행정구역 단위 마케팅에서 탈피하여 군집 특성에 맞는 마케팅을 시행한다면 비용절감 효과가 높을 것으로 추정된다.

V. 결론

5.1 연구결과 의의

초고속인터넷 가입자 확대를 위해 요금할인, 결합상품, 상품권지급 등 다양한 수단으로 가입자 유치 마케팅을 시행 하고 있다. 가입자 증대를 위해서는 신규가입자 확보만큼 중요한 것이 기존 가입자 해지방어이다. 하지만 언제 어디서 어느 고객이 해지신

청 할지 모르기 때문에 해지방어 보다는 가입자 유치 마케팅에 주력한다. 포화된 시장에서 한 통신사의 신규고객 유치는 타 통신사의 해지와 직결된다. 실효성 없이 뺏고 빼앗기기를 반복하는 시장 환경이 마케팅 비용의 주된 증가 요인으로 볼 수 있지만 달리 뾰족한 방안이 보이지 않는다. 경쟁상황에서 시장점유율이 마케팅부서의 성과와 연결되기 때문이다(양유석, 2006). 이러한 상황에서 지역별 해지를 예측할 수 있다면 좀더 효과적인 마케팅을 진행할 수 있을 것이라는 가정 하에 해지율에 대한 유의미한 예측 가능여부를 확인하고 해지 예측 정확도가 높은 최적의 지역 단위를 설정하고자 하였다.

본 연구에서는 행정구 단위, 전체, 군집 단위로 구분하여 ARIMA 시계열 분석을 통한 예측 오차율을 측정하였다. 연구결과 예측 오차율이 구 단위는 21.7%, 전체는 22.9%, 군집 단위는 20.7%로 3개 단위 전부 비교적 정확한 예측 수준에 가깝게 나타났다. 따라서 시계열 분석을 통한 해지율 예측이 유의미한 작업이라는 결과를 얻었으며 행정구역 단위보다 군집 단위의 예측 정확도가 높은 것으로 나타났기 때문에 마케팅 수행 시 행정구역 단위보다는 군집 단위 별 마케팅이 유효할 것이라고 추정되었다.

본 연구의 의의는 다음과 같다. 첫째, 해지율 예측이 비교적 정확한 범주에 속하는 것으로 나타나 해지율 예측작업이 유의미하다는 것을 확인하였다. 둘째, 공간정보를 기반으로 도출한 유사 군집 단위의 예측 정확도가 행정구역 단위보다 높다는 것을 확인함으로써 고객의 특성이 지역별로 다르기 때문에 현재와 같은 행정구역 단위 마케팅을 지양하고 군집 단위 지역에 적합한 마케팅을 시행한다면 비용절감은 물론 효과도 높을 것으로 추정된다.

또한, 본 연구의 학술적 의의는 최적의 해지율 예측 단위를 도출하기 위하여 공간검색 통계량을 활용한 군집분석을 시행했다는 점이다. 일반적인 분석 단위인 행정구역에 국한하지 않고 인접지역을 중심

으로 해지율이 유사한 지역들을 군집으로 형성하였고, 이렇게 도출한 군집 단위로 분석을 시도하여 행정구역 단위보다 예측 정확도가 높다는 것을 확인하였다.

5.2 연구의 한계 및 향후 연구

본 연구에서는 해지율 예측 정확도가 높은 지역 단위를 찾고자 행정구 단위, 군집 단위 그리고 전체로 구분하여 ARIMA 모형으로 분석 및 예측하였다. 연구결과 예측 값이 비교적 합리적인 예측의 범주로 나타나 마케팅에 적용가능하고, 유사군집 단위의 예측이 보다 유효하다는 의미 있는 결론을 얻었다. 그러나, 군집 단위 예측을 위해 유사 지역을 클러스터링할 경우 군집에 형성되지 않는 행정동들이 생기는데 이런 행정동들에 대한 예측은 하지 못한다. 이번 연구에서도 총 47개 동 중 클러스터링된 28개 동을 제외한 19개 동은 예측할 수 없었다. 따라서 공간검색통계를 활용한 군집 단위 예측은 지역 전체를 대상으로 하기보다는 타겟 마케팅에 적합하다고 볼 수 있다.

그리고, 해지율 외에 다양한 외생 변수들을 고려할 경우, 예측을 보다 향상시킬 수 있을 것으로 기대된다. 향후에는 가구 단위로 가입하는 초고속인터넷의 해지에 영향을 줄 수 있는 이혼률, 가구 전출률, 실업률 등의 변수들을 포함하여 VAR 다변량 분석을 수행하고자 한다. VAR 분석을 통해 해지율 예측은 물론이고 이러한 외생 변수들과 해지변수 간 인과관계분석, 외생변수들이 시간이 지남에 따라 해지변수에 어떻게 반응하는지를 나타내는 충격반응함수 등의 좀 더 상세한 분석을 통해 영향 변수를 찾아 마케팅에 활용이 가능할 것으로 보여진다.

향후에는 적합한 마케팅을 위해 좀 더 세분화한 분석이 필요하다. 본 연구에서는 해지율이 유사한 행정동으로 군집을 구성하였기 때문에 마케팅 고려 시

해지율의 높고 낮음에 따른 기획만 가능하다. 하지만, 해당 군집 지역의 연령별 분포, 세대 수, 부동산 평당 매매가, 자동차 등록 비율 등의 지역 데이터를 연계하면 적합하고 다양한 마케팅이 가능하다. 예를 들면, 학생층이 많은 경우 가입을 유지하는 조건으로 학원과 연계한 이벤트를 할 수 있고, 평당 매매금액이 낮거나 승용차 등록 비율이 낮거나 화물차 등록비율이 높은 지역의 경우 경제적인 상황을 고려하여 현금 지급 마케팅 등이 유효할 것으로 추정된다.

또한, 초고속인터넷 해지 유형을 보면 약정이 만료된 해지, 약정 중 해지, 무 약정 해지 이렇게 크게 세 가지로 나눌 수 있다. 본 연구에서는 해지 유형 구분을 하지 않고 전체 해지율로 분석 하였으나 향후에는 세 가지 유형을 각각 나누어서 분석하면 해지 예측 정확도가 향상 될 수 있을 것으로 기대된다.

REFERENCES

- Assessment of Telecommunication Market Competition(2018), Korea Information Society Development Institute. [printed in Korean]
- Baek. S. H.(2006), Study on air travel demand models and elasticities for Jeju-mainland domestic routes, master's thesis, University of Seoul, Seoul. [printed in Korean]
- Bambhroliya AB, Burau KD and Sexton K.(2012), "Spatial Analysis of County-level Breast Cancer Mortality in Texas," *Journal of Environmental and Public Health*, A ID 959343, 8.
- Cho. B. S. and Ko.j.g(2002), "A Study on Telcos Strategy and Forecasting of Post ADSL," *Korea Institute Of Communication Sciences*, 109-112. [printed in Korean]
- Choo. G. S.(2003), An application of Bass Diffusion Model to Wire · Wireless Internet Service Demand Forecasting, master's thesis, University of Chung-ang, Seoul. [printed in Korean]
- Cho. S. S and Jeong. D. S.(2002), "The Comparative Evaluations of Telecommunications Service Forecasting Models for Forecasting Performance," *Electronics and Telecommunications Trends*, vol.17. [printed in Korean]
- Dahly D. Obesity clustering in Cebu(2011), "Philippines: an application of satscan and the spatial scan statistic," *Journal of Epidemiology and Community Health*, 65, A71
- Han, J. H., Lee, M. J.(2016), "Cancer cluster detection using scan statistic," *Korean Data Information Science Society*, vol.27(5) 1193-1201. [printed in Korean]
- Jung I, Kulldorff M, Klassen A.(2007), "A spatial scan statistic for ordinal data. *Statistics in Medicine*," 26 , 1594-1607
- Jung I, Kulldorff M, Richard OJ.(2010), "A spatial scan statistic for multinomial data. *Statistics in Medicine*," epub
- Kim, B. S.(2014), "Forecasting Passenger Transport Demand Using Seasonal ARIMA Model," *Korean Society For Railway*, vol.17(4) pp. 307-312. [printed in Korean]
- Kim. J. G.(2011), study on forecasting container traffic of port using SD and ARIMA, master's thesis, University of Incheon. [printed in Korean]
- Kim. H. B.(2017), Hybrid Model of Time-series and Internet Search Queries for Short-term Forecasting of Housing Transaction Volumes, master's thesis, University of Inha, Incheon. [printed in Korean]
- Kim. K. Y., Kim. H. S.(2011), "KTX Passenger Demand Forecast with Intervention ARIMA

- Model," *Korean Society for RailWay*, 14(5), 470-476. [printed in Korean]
- Korea Internet & Security Agency(2018), "2018 Internet Usage Survey". [printed in Korean]
- Lee. K. J, Hwang. M. H, Han. S. H and Yang. E. J. (2015), *First Steps to Understanding and Utilizing Spatial Statistics Analysis*, Korea Research Institute for Human Settlements. [printed in Korean]
- Lee, S. H., Kim, H. C. and Park, M. S.(2016), "The Effects of Free Offers in Bundled Products: A Case of Korean Broadcasting and Telecommunication Market," *Korea Academic Society of Industrial Organization*, vol.24 (1). [printed in Korean]
- Nam. J. O, Noh. S. G. and Park E.Y.(2012), "A Study on Forecasting Oyster Price with Time Series Analysis," *Korean Society of Oceanography*. [printed in Korean]
- Noh. Y. J.(2016), "A Comparison Study on Statistical Modeling Methods," *Korea Academia-Industrial cooperation Society*, 17(5), 645-652. [printed in Korean]
- Park. C. H and Hong. J. H.(2009), "Market Status and Pricing Policy under Regulatory Environment," *Korea Business Review*, vol 13(1). [printed in Korean]
- Park. J. H, Kim. W. J, Park. S. U.(2017), "A Study on the Regional Characteristics of Broadband Internet Termination by Coupling Type using Spatial Information based Clustering," *Intelligence and Information System*, vol23(3). [printed in Korean]
- Park. J. O, I. S. Choi and M. H. Na(2016), "A Study on the effects of air pollution on circulatory health using spatial data," *J Korea Soc Qual Manag*, Vol.44(33), 677-688. [printed in Korean]
- Pervaiz F, Pervaiz M, Abdur Rehman N, Saif U. (2012), "early epidemic detection from Google flu trends," *Journal of Medical Internet Research*, [J Med Internet Res], Vol.14
- Green C, Hoppa RD, Young TK, Blanchard JF. (2003), "Geographic analysis of diabetes prevalence in an urban area," *Social Science and Medicine*, 57:551-560, 2003.
- Woo. K. S. and Lee. S. S.(2014), "Time Series Modeling for Forecasting Land Price Change Rate," Korea Lnad Corporation. [printed in Korean]
- Woo. K.(2002), "Forecasting the Fluctuation Rate of Land Price with Intervention ARIMA in Korea land Market," *Korea Research Institute for Human Settlements*, vol35, 51-64. [printed in Korean]
- Yang. Y. S and Kim. Y. K.(2006), "Managing Market Share in the Regulated Telecommunication Market," *Korea Business Review*, 10(1)
- Yoo. C. S., Lee. J. H, Yang. D. M. and Jeong. J. H. (2011), "Spatial Analysis of Rain Gauge Networks," *Korean Society of Hazard Mitigation*, 11(4). 179-187. [printed in Korean]
- Yoon. H. Y and Park. S. S.(2019). "Analysis and Estimation of Food and Beverage Sales at Incheon Int'l Airport by ARIMA-Intervention Time Series Model," *Korea Academia-Industrial*, 20(2), 458-468. [printed in Korean]
- Yoon. J. H. and Kim. H. S.(2000), "Short-term Forecasting Model for the Housing Market," Korea Research Institute for Human Settlements. [printed in Korean]
- Yoon. J. S, Huh. N. K, Kim, S. Y. and Huh. H. K. (2010), A Study on International Passenger and Freight Forecasting Using the Seasonal Multivariate Time Series Models, *Korean Statistical Society*, 17(3), 473-481. [printed in Korean]

국내참고문헌

- 김관형, 김한수(2011), “개입 ARIMA 모델을 이용한 KTX 수요예측,” **한국철도학회**, 14(5).
- 김범승(2014), “계절ARIMA모형을 이용한 여객수송수요 예측,” **한국철도학회**, 17(4), 307-312.
- 김종길(2011), “SD기법과 ARIMA를 적용한 컨테이너물동량 예측에 관한 연구,” 물류경영학 석사학위논문, 인천대학교 동북아물류대학원.
- 김한빈(2017), “주택거래량 단기 예측을 위한 시계열과 인터넷 검색어의 혼합모형,” 토목공학 석사학위논문, 인하대학교.
- 남중오, 노승국, 박은영(2012), “시계열 분석을 이용한 굴 가격 예측에 관한 연구,” **해양정책연구**.
- 노유정(2016), “통계모델링 방법의 비교 연구,” **한국산학기술학회**, 17(5), 645-652.
- 박찬희, 홍진환(2009), “규제적 환경에서의 시장지위와 가격정책,” **Korea Business Review**, 13(1).
- 박장혁, 김우주, 박상언(2017), “공간정보기반 클러스터링을 이용한 초고속인터넷 결합유형별 해지의 지역별 특성연구,” **한국지능정보시스템학회**.
- 박진욱, 최일수, 나명환(2016), “공간 자료를 이용한 대기 오염이 순환기계 건강에 미치는 영향 분,” **한국품질경영학회**.
- 백승한(2006), “제주-내륙 간 국내선 항공여객수요모형 및 탄력성에 관한 연구,” 환경계획 석사학위논문, 서울대학교.
- 양유석, 김영곤(2006), “정부규제와 시장점유율,” **Korea Business Review**, 10(1).
- 우경, 이성석(2014), “지가변동률 예측을 위한 시계열 모형 분석,” 한국토지공사.
- 우경(2002), “개입-ARIMA 모형을 이용한 지가변동 예측에 관한 연구,” **국토연구**, 35, 51-64.
- 유철상, 이지호, 양동민, 정재학(2011), “유량계망의 공간 분석 : 일양 분포와 포아송 분포의 적용,” **한국방재학회**, 11(4).
- 윤주현, 김혜승(2000), “주택시장 경기동향 및 단기전망 연구,” 국토연구원.
- 윤지성, 허남균, 김삼용, 허의영(2010), “계절형다변량시계열모형을이용한국제항공 여객및 화물 수요예측에 관한 연구,” **한국통계학회**, 17(3), 473-481.
- 윤한영, 박성식(2019), “시계열 모형을 이용한 인천국제공항 식음료 매출 분석 및 추정 연구,” **한국산학기술학회**, 20(2), 458-468.
- 이경주, 황명화, 한선희, 양은정(2015), “공간통계 분석의 이해와 활용을 위한 첫걸음,” 국토연구원.
- 이상현, 김현철, 박민수(2016), “방송통신서비스 결합상품에서의 공짜마케팅효과,” **산업조직연구**.
- 조병선, 고중걸(2002), “통신사업자들의 Post ADSL구축 전략 및 가입자 예측,” 한국통신학회 학술대회 논문집.
- 조상섭, 정동진(2002), “통신서비스산업 예측모형 예측력 비교 분석,” **전자통신동향분석**, 17.
- 추길수(2003), “Bass확산모형을 이용한 유.무선인터넷 서비스의 수요예측,” 경제학과 석사학위논문, 중앙대학교.
- 한준희, 이민정(2016), “스캔통계량을 이용한 암 클러스터 탐색,” **한국데이터정보과학회**.
- 과학기술정보통신부, 인터넷진흥원(2018), “인터넷 이용실태 조사보고서”.
- 정보통신정책연구원(2015), “통신시장 경쟁상황 평가(2015년도)”.

A Study on Improving Accuracy of Predicting Broadband Internet Termination Rate by Using Spatial Clustering

Jang-hyuk Park* · Sang-un Park** · Woo-ju Kim***

Abstract

The marketing battle among Korean telecom companies to secure new subscribers has been so fierce that their marketing expense now exceeds operating profit by more than double. The cost of this battle, however, is excessive since it is nothing but a zero-sum game, which each participant's gain or loss of utility is exactly balanced by the losses or gains of the utility of the other participants. The general marketing strategy of telecom companies is not effective since the companies mainly focus on taking other companies' customers and do not consider maintaining own subscribers. By increasing the accuracy of measuring termination rate of existing broadband internet users, telecom companies can be benefited saving tremendous amount of money.

The purpose of this study is to identify the possibility of forecasting the Broadband internet termination rate by using time series analysis and to set the optimal region units that have condition in which the termination rate can be accurately predicted. The 63,289 termination data of A company were collected from 47 units of administrative dong, which is similar concept to town U.S., in city of Anyang, Kunpo and Uiwang. These are located in kyonggi-do, Korea. The unit for prediction and verification of termination rate were divided by administrative districts and spatial clustered dong, and measuring accurate termination rate was better in order of cluster, district, and total unit. Overall, the results revealed that extracting data from spatial cluster is significant in predicting the termination rate. This shows that it may be effective to predict and implement the termination rate in unit of cluster when telecom companies use marketing strategy in different regions.

Key Words: ARIMA, Broadband Internet, Predict Rate of Termination, Spatial Clustering, Time Series Analysis

* Ph.D. Candidate, Technology Policy, yonsei University, First Author

** Professor, Department of Management Information Systems, Kyonggi University, Corresponding Author

*** Professor, Department of Information and Industrial Engineering, Yonsei University, Co-author