Journal of KIIT. Vol. 12, No. 10, pp. 113-121, Oct. 31, 2014. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/kitr.2014.12.10.113

공공데이터를 이용한 G2B 전자상거래 시장수요예측 - 조달청 사례를 중심으로 -

박현기*, 안재경**

Demand Forecasting for G2B E-commerce Using Public Data : A Case Study of Public Procurement Service

Hyunki Park*, Jaekyoung Ahn**

본 논문은 서울과학기술대학교 교내 연구비 지원으로 수행되었슴.

요 약

본 논문은 조달청 종합쇼핑몰 공공데이터를 체계적으로 활용하여 ARIMA 모형 기반의 시장수요예측 프로그램을 구현하였다. 최근 공공데이터의 활용과 연구에 대한 관심이 높아지고 있으나, 공공데이터의 수집, 가공 및 분석을 체계적으로 수행한 연구는 제한적인 실정이다. 공공데이터를 활용하여 분석하기 위해서는 관련데이터의 속성을 구체적으로 파악하여 사용자의 의도에 맞도록 데이터를 추출하여야 하며, 이를 토대로 연구하고자 하는 방법론을 적용하여야 하지만, 현재까지의 연구에서는 이를 처리하는 데 한계가 있었다. 본 연구는 이러한 한계점을 보완하고자 공공에서 제공하는 데이터의 속성을 파악하고 추출하였으며, 프로그래밍을 통하여 일괄적으로 부문별 시장규모를 예측할 수 있는 방법을 제시하였다. ARIMA 모형의 단위근 검정, 모형식별, 모형추정 및 모형검증에 대한 전 과정을 프로그래밍화하여 일괄적으로 처리하고 예측함으로써 공공데이터의 처리와 분석이 가능함을 제시하였다. 분석결과 MAPE 값이 3.90%~ 24.47%로 본 연구에서 구현한 ARIMA 모형을 이용한 처리방법이 우수한 것으로 나타났다.

Abstract

In this paper, an ARIMA based demand forecasting program has been implemented by utilizing public data on G2B e-commerce at public procurement service. Recently, there has been growing interest in research and use of public data. However, it is found that few studies proposed systematic procedures for collecting, processing, and analysing the public data. There also have been limitations on grasping the attributes of the public data, extracting them for researchers' purpose and applying the right methodologies. This study presents a series of procedures for redefining and extracting the public data, and attempts to forecast the market sizes in each segment using the ARIMA model. This paper also sheds light on utilizing the public data in the area of demand forecasting by programming the whole processes where unit root test, model identification, model estimation, and model diagnostic checking are automatically to be done. As a result of the analysis, MAPE values fall into 3.90%～24.47%, which shows that the ARIMA model implemented in this paper is working well.

Keywords
public data, G2B, R, ARIMA, demand forecast, time series analysis

* 서울과학기술대학교 IT정책전문대학원 산업정보시스템공학과
** 서울과학기술대학교 글로벌융합산업공학과 교수 (교신저자)
· 접 수 일: 2014년 08월 14일
· 수정완료일: 2014년 09월 29일
· 게재확정일: 2014년 10월 01일
· Received: Aug. 14, 2014, Revised: Sep. 29, 2014, Accepted: Oct. 01, 2014
· Corresponding author: Jaekyoung Ahn
Department of Industrial & Information Systems Engineering, Seoultech, 139-743, 232 Gongneung-Ro, Nowon-Gu, Seoul, Korea,
Tel.: +82 51-629-5740, Email: jkahn@seoultech.ac.kr

Ⅰ. 서 론

2013년에 제정된 “공공데이터의 제공 및 이용 활성화에 관한 법률”과 정부의 공공데이터 개방 정책에 따라 각 공공기관이 폐쇄적으로 관리하고 운영하였던 데이터를 민간에 개방하기 시작하였다(www.data.go.kr). 공공데이터는 오픈 API 형태나 EXCEL 파일 등 다양한 형태로 제공되며, 민원요청에 의해서도 제공되고 있다. 또한 국책연구기관에서 진행된 연구 데이터까지 민간에 제공되고 있어 그 양이 방대하다. 이렇게 방대한 양의 공공데이터는 최근 이슈가 되는 빅데이터의 처리와 함께 관심이 높아지고 있다[1]. 빅데이터 분석은 한순간에 나타난 현상이 아니며 일반기업 및 공공기관에서는 인터넷의 발달과 데이터의 디지털화 이후부터 지속적으로 발달되어왔다.

현재 일반기업의 마케팅 활동에서도 빅데이터의 활용을 쉽게 찾아볼 수 있으며, 전자상거래 모델 분석에서 주로 쓰이고 있다. Yahoo, Google, eBay, Amazon과 같은 외국 기업에서는 웹 기술의 진화에 맞춰 검색엔진 기능을 확장하여 기업과 고객을 직접 연결하는 비즈니스 모델을 수립하였고, 이를 통해 기업과 고객과의 상호작용을 극대화하여 다수의 기업과 고객을 확보하는 데 사용하고 있다[2]. 국내 기업의 경우 전자상거래에서 수집된 데이터를 고객의 마케팅에 활용하고 있으며, 관련연구도 증가하고 있다. 하지만 기존의 연구들에서는 설문조사, 기업에서 일부 공개한 데이터와 정부에서 공개한 데이터에만 의존하여 분석하고 있다[3].

최근 정부의 공공데이터 개방이 확대되어 그동안의 연구와 달리 다양한 연구가 가능해짐에 따라 공공 부분, 관련 기업, 연구기관 등에서는 이를 적극적으로 활용하는 방안을 제안하고 있다[4]. 하지만 기존 연구에서는 통계청에서 제공된 정형화된 데이터에 대한 분석과 활용에 그치고 있으며, G2B 전자상거래 규모에 대한 포괄적인 수요예측모형을 연구했기 때문에 각각의 분류별 세부시장을 예측하는 데는 한계가 있다. 더욱이 이러한 데이터를 활용하기 위해서는 각 데이터를 연구의 목적에 적합하도록 가공하는 방법에 대한 연구가 필요하다. 따라서 본 연구는 공공데이터의 수집, 가공 및 분석에 대한 방법을 제시하고 수집된 데이터를 이용하여 수요예측을 수행하고자 한다. 본 연구에서는 조달청 나라장터의 데이터를 이용하여 국내 G2B 시장규모를 국제표준 분류코드에 따라 세분화하여 예측한다. 이를 위해 R 프로그램을 이용하며, Box-Jenkins의 ARIMA(Autoregressive Integrated Moving Average)모형을 적용하고자 한다[5].

본 연구는 R 프로그램을 이용해 ARIMA 모형을 프로그래밍 하였으며, 분석결과 MAPE(Mean Absolute Percent Error)의 값이 3.90%~ 24.47%로 나타나 본 연구에서 구현한 ARIMA 모형의 처리방법이 우수한 모형임을 입증하였다.

Ⅱ. 관련 연구

빅데이터는 대규모, 현실성, 시계열성, 결합성 등과 같은 특징을 가지고 있으며, 종류로는 정형, 반정형, 비정형 등으로 나눌 수 있다. 이러한 데이터의 종류는 분석자와 분석방법에 따라 다양한 해석이 가능하다[6]. 이중 시계열성을 가진 데이터는 과거 데이터를 학습하여 미래의 데이터를 예측할 수 있는 데이터이다.

시계열 데이터를 이용한 연구는 어획량분석, 기상변화, 가정용 전기소비와 같은 현실적인 분석부터 저가항공 여객운항 수요예측, 국내 GDP 예측, 전자상거래 예측 등과 같이 광범위하게 이루어지고 있다. 계절적인 요인을 고려한 연구로는 어획량 데이터를 분석하여 이중모형을 산출하고 그것을 예측하는 방법을 제안한 연구와 댐의 유입량을 예측하기 위해 24년 동안의 연도별 유입량을 이용하여 월 평균 유입량을 예측하는 연구 등이 존재한다[7][8].

R 프로그램을 이용한 연구로 가정용 전기소비량의 월, 분기, 년과 같이 다양한 시계열형태의 데이터를 변환하고 최적의 예측방법을 찾아 적용한 것이 있다[9]. 또한 ARIMA 모형의 모수 추정을 최소화하여 이를 저가항공의 여객운항에 대한 수요예측을 적용한 연구가 있다[10]. 국내 시장규모예측에 관한 연구는 공공데이터 개방 이전부터 수행되어왔으며, 이러한 연구에서 사용된 대표적인 데이터는 통계청에서 제공하는 데이터이다. 관련 연구로는 국내총생산(GDP)의 예측을 위해 분기별 데이터를 이

용하여 예측한 연구[11]와 전자상거래의 월별과 분기별 데이터를 이용하여 로지스틱 모형 등 다양한 예측방법을 비교하고 B2C, B2B, B2G 부분에 ARIMA 모형을 적용한 것이 있다[12].

정부 3.0 시대의 공공데이터 개방은 광범위한 레벨의 데이터를 활용하여 보다 다양한 분석을 하는데 그 목적이 있으며, 더 나아가서는 이를 정책에 반영하여 국가적인 효율성을 극대화하는 데 있다[13]. 전자상거래와 관련한 기존 연구에서 사용된 데이터는 통계청에서 제공하는 정형화된 데이터가 사용된다. 하지만 통계청에서 제공하는 데이터는 국내 G2B의 전체 매출액으로 전체 시장규모를 예측하는 데는 적절하지만 국제적으로 세분화되어 있는 G2B 전자상거래의 부문별 시장규모를 예측하고 활용하는 데는 한계점이 있다[12]. 즉 기존 연구에서 사용된 데이터는 정형화된 데이터로서 전체 시장규모예측은 가능하지만 세부적인 분석에는 한계가 있다. 정형화된 데이터는 분석이 용이한 장점이 있지만 연구자의 연구목적에 맞는 데이터 가공이 어렵다. 연구목적에 맞는 분석을 위해서는 데이터 자체의 속성과 정보를 고려해야하기 때문이다. 따라서 본 연구에서는 이러한 전자상거래 부분 중 국가와 기업 간의 시장 규모를 예측하고 대용량 데이터의 처리에 내재하고 있는 한계점을 개선하고자 한다. 이를 위해 G2B 전자상거래 부분의 공공데이터를 수집 - 가공 - 분석하고 이에 대한 수요예측 프로그래밍 방법을 제안하고자 한다.

III. 데이터 가공 · 분석 방법

3.1 데이터 수집·가공

국내 G2B 전자상거래의 대표적인 공공기관인 조달청 사이트에 개방된 나라장터 쇼핑몰 판매실적 데이터를 조회한 결과, 연구 목적과 유사한 데이터를 제공하는 것을 확인하였다[1]. 그러나 기본적으로 제공하는 데이터만으로는 수요예측이 어렵기 때문에 정부의 공공데이터 포털사이트를 통하여 조달청에 데이터 공개신청을 하였다. 연구에 필요한 데이터를 EXCEL 파일로 다운로드 받을 수 있게 시스템 수정을 요청하였으나 공개되는 데이터를 일괄적으로 받기에는 데이터의 양이 방대하고 연구에 사용할 데이터를 전부 시스템에서 조회하기에는 시스템 용량의 한계가 있어 연월로 데이터를 조회할 수 있도록 표 1과 같이 수정하여 요청하였다. 데이터는 2000년 1월부터 2014년 6월까지 월 단위 EXCEL 파일로 다운로드 하였고, 데이터 구조화 시 연월 데이터 필드를 추가하여 가공하였다.

표 1. 데이터 필드 비교표
Table 1. Data fields comparison table

구 분	데이터 필드 내용
개방 데이터	물품식별번호,물품식별명,조달업체명,수량합계,평균단가,증감금액합계
요청 후 제공 데이터	**물품대분류번호, 물품대분류명, 물품중분류번호, 물품중분류명, 물품소분류번호, 물품소분류명, 물품분류번호, 물품분류명,** 물품식별번호, 물품식별명, 조달업체명, 수량합계, 평균단가, 증감금액합계

3.2 데이터 구조화 · 데이터베이스화

시계열 데이터 분석을 위해서는 사용되는 데이터가 연속성을 가지고 있어야 한다. 데이터 구조화 및 데이터베이스화를 통하여 데이터의 연속성을 확보하였고, 연구에 사용할 데이터의 추출과 검증은 SQL(Structured Query Language)문을 사용하였다.

데이터 속성을 분석한 결과 판매실적 데이터는 총 4단계의 계층적 구조로 되어 있으며, 4단계의 분류 기준은 물품목록정보에 법률에 따라 UN에서 제정한 국제표준 분류코드(UNSPSC : The United Nations Standard Products and Services Code)를 기반으로 표 2와 같이 정하고 있다. 전자상거래 쇼핑몰인 나라장터에서는 이러한 표준에 의해 시스템을 구축하고 지속적인 관리를 하고 있다[14][15].

각 단계는 2자리의 숫자 값과 텍스트로 구성되어 있으며, 각 테이블 필드는 제공된 데이터의 속성을 분석하여 표 3과 같이 설계하였다. 데이터베이스 시스템은 SQL 서버 2005버전을 이용하였으며, 제공된 EXCEL 데이터를 데이터베이스화하기 위한 데이터 마이그레이션 작업은 DTS(Data Transformation Services)엔진을 이용하여 설계된 테이블로 수행하였다.

표 2. UNSPSC 구조
Table 2. UNSPSC structure

단계	단계명	분류명	설 명
1	세그먼트 (Segment)	대분류	분석 목적을 위한 논리적인 집단
2	패밀리 (Family)	중분류	일반적으로 인정되는 상호관계 상품그룹
3	클래스 (Class)	소분류	공통적인 특징을 가지고 있는 그룹
4	커머디티 (Commodity)	분류번호	3단계의 특징을 가진 제품과 서비스

표 3. 테이블 필드 설명
Table 3. Table field description

필드명	데이터타입	설 명
MONTH_MM	char(6)	연월
LEVEL1	char(2)	세그먼트번호
LEVEL1_NAME	varchar(510)	세그먼트명
LEVEL2	char(4)	패밀리번호
LEVEL2_NAME	varchar(510)	패밀리명
LEVEL3	char(6)	클래스번호
LEVEL3_NAME	varchar(510)	클래스명
LEVEL4	char(8)	커머디티번호
LEVEL4_NAME	varchar(510)	커머디티명
PRODUCT_CD	varchar(8)	물품식별번호
SALE_SUM_EK	decimal(18,0)	증감금액합계

3.3 데이터 추출 · 분석대상 선정

데이터의 추출과 분석을 위해 SQL(Structured Query Language) 문을 사용하여 데이터를 추출하고 분석하였다. 분석결과, 나라장터 쇼핑몰에서 사용된 분류는 1단계(세그먼트) 분류가 50개, 2단계(패밀리) 분류가 215개, 3단계(클래스) 분류가 745개, 4단계(커머디티) 분류가 2,554개를 사용하고 있으며, 전체 데이터는 총 3,563,755건이다.

UNSPSC 구조에 따라 분석 목적을 위한 논리적인 집단인 1단계(세그먼트) 분류를 분석대상으로 하였고, 실험에 사용하고자 하는 세그먼트 분류는 연속적으로 판매실적이 있는 분류만 추출하였다. 1단계 세그먼트에서 월별 데이터가 없는 데이터를 제외한 21개의 분류 중 데이터 자체 분류가 모호한 기타 분류는 선정에서 제외하여 20개의 분류를 표 4와 같이 분석대상으로 선정하였다.

표 4. 선정된 분류
Table 4. Selected categories

코드	세그먼트명
11	광물,직물및비식용동식물자원
12	화학제품
14	종이원료및종이제품
21	농.수.임.축산용기계
25	상용,군용,개인용운송기구및액세서리와부품
26	회전기기및경전기
30	건자재
31	제조부품
39	전기시스템,조명,부품,액세서리및보조용품
40	배관유체조절시스템장비및부품
43	정보기술방송및통신기
44	사무용기기액세서리및용품
45	인쇄,사진및시청각기기
46	공공안전및치안장비
47	위생장비및용품
52	가정용품및가전제품
53	의류,가방및개인관리용품
55	출판물
56	가구및관련제품
60	악기,게임,장난감,미술작품,교육용품및보조품 등

그러나 월별 데이터는 국내의 기후변화, 명절, 물류 파업 등과 같은 불규칙요인이 포함되어 시장규모를 예측하는 데는 한계점이 있기 때문에, 통계청에서 발표하는 시장규모 및 산업동향을 예측하는 경우에도 월별 예측보다는 분기별 예측을 통하여 지표들을 발표하고 있다. 따라서 본 연구는 이러한 불규칙 요인을 제거하고자 앞선 선행된 연구와 같이 월별 데이터를 분기별로 합산하여 데이터를 동일한 기준으로 가공하였다[12][16][17]. 예측을 위해 2000년 1/4분기부터 2013년 4/4분까지 학습데이터로 사용하였고 2014년 1/4분기와 2/4분기의 데이터는 예측 값을 비교하는데 사용하였다.

3.4 데이터 예측방법 설계

수요예측을 위해 사용할 예측방법은 그림 2와 같이 ARIMA 모형을 이용하여 설계하였다. 시계열 데이터를 ARIMA 모형에 적용하기 위해서는 단위근 검정, 모형식별/추정, 모형검증, 예측과 같이 4단계 과정을 거쳐야 한다. 대부분의 선행연구[17]에서는

4단계를 개별적인 절차를 거쳐 분석하였으나, 본 연구에서는 R 프로그램에서 제공하는 라이브러리 함수와 프로그래밍을 통하여 20개 분류에 대한 단위근 검정, 모형식별, 모형검증 등을 그림 2와 같이 일괄 처리한다.

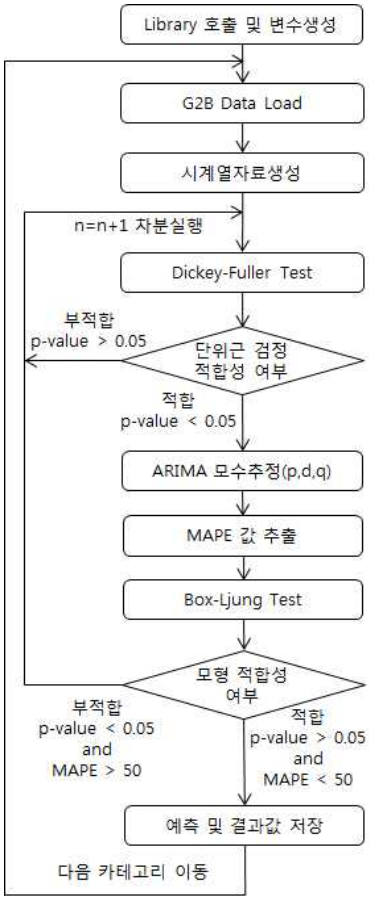

그림 2. 데이터 분석 순서도
Fig. 2. Data analysis flowchart

IV. 데이터 분석과정 및 결과

4.1 데이터 배열 및 변수 설명

각 분류별 수요예측을 일괄적으로 처리하기 위해서는 데이터의 배열과 처리를 위한 변수 설명이 필요하다. 데이터의 배열은 분석을 위한 시계열 데이터를 임시로 저장하고 처리하는 역할을 하며 표 5와 같이 정의한다.

표 5. 배열 설명
Table 5. Array description

배열명	설 명
G2B_DATA_FRAME	원시데이터 정보를 데이터프레임
arrayDataGroup	모형분석용 데이터 요약정보
arraySqlText	분석데이터 추출을 위한 SQL문 정보
arrayTimeData	시계열 타입의 분석데이터 정보
arrayPredictData	ARIMA모형을 통한 예측정보
arrayAdfData	단위근 검정 데이터 정보
arrayArimaPDQ	ARIMA모형의 모수정보
arrayArimaInfo	그룹별, ARIMA(p,d,q), 예측결과 정보

표 6. 변수 설명
Table 6. Parameter description

변수명	설 명
intArimaDiff	모형의 차분 값
intArimaAic	최종 AIC 결과값
intTempAic	비교를 위해 AIC값을 저장하는 변수
adfTestResult	단위근 검정의 결과
Mape	MAPE 값의 결과
boxTestResult	Ljung-Box Test의 결과
arimaResult	ARIMA모형의 최종 결과

사용된 변수는 사용자의 의사결정을 프로그래밍하는데 필요하며, 표 6과 같이 설정한다. 변수들은 ARIMA 모형의 모수인 차분 값(d값)에 대한 결정과 과대적합 진단을 통한 p, q값을 결정하는 데 사용하고, 모형의 진단인 Ljung-Box Test를 검증하는 데 사용한다.

4.2 프로그래밍 주요설명

시계열 데이터 생성을 제외한 분석의 전체 과정은 그림 3과 같이 프로그래밍 하였다. 배열의 ① for문은 arrayDataGroup 데이터 배열의 길이인 구분자 코드만큼 반복문을 사용하여 시계열 데이터를 생성하는 것이다. 배열 ②의 for문은 단위근 검정의 차분 값을 결정하고, ③의 for문은 과대적합 진단법으로 모수를 추정한다. ④Ljung -Box와 MAPE 테스트로 모형의 적합성을 판단한다.

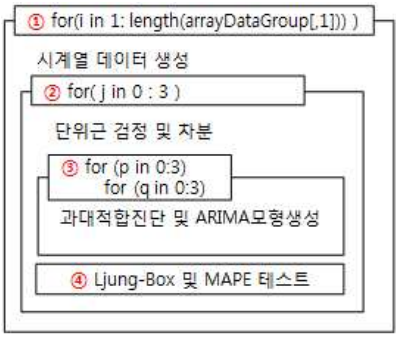

그림 3. R 프로그래밍 배열 구성도
Fig. 3. R programming array configuration

4.2.1 시계열 데이터 생성

원시데이터 정보를 분석에 적합한 시계열 데이터로 생성하기 위해서는 R 프로그램의 데이터프레임 구조를 사용하여야 한다. 데이터베이스에서 가공된 정보를 CSV(Comma-Separated Values)파일로 생성하고, R프로그램으로 읽는다. 라이브러리는 데이터프레임 구조의 데이터 집합이며 데이터베이스에서 사용하는 SQL문처럼 R프로그램에서 사용하게 할 수 있는 sqldf 라이브러리를 사용한다. 일괄적으로 처리하기 위해 SQL문을 만들어 리스트 형식으로 저장하고 sqldf함수를 사용하여 데이터를 추출하게 프로그래밍 하여 그림 4와 같이 시계열 데이터를 생성한다.

```
# 데이터 그룹화
arrayDataGroup <- sqldf(
        "select  gubun_cd
        ,min(year_mm) as min_year_mm
        ,max(year_mm) as max_year_mm
        ,min(substr(year_mm,1,4)) as start_year
        ,min(substr(year_mm,5,2)) as start_mm
        ,count(*) as data_length
        from G2B_DATA_FRAME
        group by gubun_cd")

# SQL문 생성
arraySqlText[[i]] <- paste(
         " select sum_ek
           from G2B_DATA_FRAME
           where gubun_cd =",arrayDataGroup[i,1],
         " and year_mm between ",arrayDataGroup[i,2],
         " and ",arrayDataGroup[i,3],
         " order by year_mm asc" )
# 시계열 데이터 생성
arrayTimeData <- ts(sqldf(arraySqlText[[i]])
       ,start=c(as.numeric(arrayDataGroup[i,4]),
                as.numeric(arrayDataGroup[i,5])),
       frequency=4)
```

그림 4. 시계열 데이터 생성 R 프로그래밍
Fig. 4. R programming for time series data generation

4.2.2 모수추정 및 모형검증

ARIMA(p,d,q)모형의 모수는 AR(p)모형과 MA(q)모형에서의 p, q값과 차분을 통하여 d값을 구한다. 본 연구에서는 모형의 모수를 추정하기 위해서 단위근 검정을 통하여 데이터의 정상과 모수인 d의 차분을 adfTestResult의 p-value 값이 0.05 미만인 데이터에 대해서만 모수를 추정하게 구현한다.

```
for( j in 0 : 3 ){
    # 단위근 검정
    adfTestResult<-adf.test(arrayAdfData)
    # 단위근 적합성 검사
    if(adfTestResult$p.value < 0.05){
        intArimaDiff <- j
        adfTestPvalue <- adfTestResult$p.value
        arrayArimaPDQ <- c(0,0,0)
        intArimaAic<-Inf
        logYN <<- 0
        # 아카이케 정보판단기준을 적용한 과대적합모형
        for (p in 0:3) {
            for (q in 0:3) {
                intTempAic <- tryCatch( {
                              AIC(arima(x=arrayTimeData
                              ,order=c(p,intArimaDiff,q)
                              ,method="ML"))
                          } , error = function(cond){
                              return(AIC(arima(x=log(arrayTimeData)
                                       ,order=c(p,intArimaDiff,q)
                                       ,method="ML")))
                          })
                if (intTempAic < intArimaAic) {
                    arrayArimaPDQ <-c(p,intArimaDiff,q)
                    arimaResult <- tryCatch( {
                                arima(arrayTimeData
                                       ,order=arrayArimaPDQ
                                       ,method="ML")
                              },error = function(cond){
                                  logYN <<- 1
                                  return(arima(log(arrayTimeData)
                                       ,order=arrayArimaPDQ
                                       ,method="ML"))
                              })
                    intArimaAic <- intTempAic
                }
            }
        }
        # MAPE값 추출
        Mape <- accuracy(arimaResult)
        # Ljung-Box 테스트
        boxTestResult <- Box.test(arimaResult$residuals,
        lag=arrayDataGroup[i,6]-arrayArimaPDQ[1]-arrayArimaPDQ[3]-1,
        type="Ljung-Box")
        # Ljung-Box && MAPE 적합성 검사
        if(boxTestResult$p.value > 0.05 &&  Mape[5] < 50){
            arrayPredictData <- predict(arimaResult, n.ahead=2)$pred
            if(logYN < 1){
                newRow <- data.frame(gubun_cd=arrayDataGroup[i,1],
                        p=arrayArimaPDQ[1],
                        d=arrayArimaPDQ[2],
                        q=arrayArimaPDQ[3],
                        Mape=as.numeric(Mape[5]),
                        MonthPred1=arrayPredictData[1],
                        MonthPred2=arrayPredictData[2])
            }else{
                newRow <- data.frame(gubun_cd=arrayDataGroup[i,1],
                        p=arrayArimaPDQ[1],
                        d=arrayArimaPDQ[2],
                        q=arrayArimaPDQ[3],
                        Mape=as.numeric(Mape[5]),
                        MonthPred1=exp(arrayPredictData[1]),
                        MonthPred2=exp(arrayPredictData[2]))
            }
            arrayArimaInfo <- rbind(arrayArimaInfo,newRow)
            break
        }
    }

    if(j > 0){
        arrayAdfData <- diff(arrayAdfData,differences=j)
    }
}
```


그림 5. 모수추정 및 모형검증 R 프로그래밍
Fig. 5. R programming parameter estimation and model validation

모수 p와 q값은 arima함수에서 제공하는 AIC (Akaike's Information Criterion)값을 비교하여 추정하는 방법을 사용하고, 데이터의 금액이 연산범위를 벗어나는 값에 대해서는 log함수를 이용하여 값을 작게 만들어 연산하게 한다. p와 q값을 이중 배열하여 가장 작은 AIC값을 찾는 과대적합법을 그림 5와 같이 프로그래밍 하였다[17][18].

Ljung-Box Test를 통하여 모형검증을 하였고 검증에서 사용된 자유도의 결정 수식은 기존의 선행연구에서 사용된 x^2-분포를 따르는 $k-p-q-1$ 수식을 사용한다[17]. k는 분석에 사용된 데이터의 수를 말하며 배열 정보(arrayDataGroup[i,6])에 저장된 데이터의 수를 사용하며, arrayArimaPDQ[1]에서 p값의 정보와 arrayArimaPDQ[3]에서 q값의 정보를 사용하여 자유도 값을 결정한다.

기존의 선행연구에서는 p-value 값만으로 모형검증을 하였으나 이는 예측값의 정확성을 판단하는 기준으로는 한계가 있어 일반적으로 평가기준에 사용하는 MAPE 값을 추가하여 모형검증을 보완한다. MAPE의 판정기준은 0%~10%는 매우 정확한 예측, 10%~20%는 비교적 정확한 예측, 20%~50% 비교적 합리적 예측으로 사용되고 50% 이상의 값은 부정확한 예측이므로 사용하지 않는다[17][19].

모형검정으로 Ljung-Box Test의 p-value 값이 0.05 보다 큰 경우와 MAPE 값이 50%로 미만인 경우 해당 분류의 모형이 정상적으로 추정이 되었다는 것을 의미한다. 해당 분류의 추정된 모수를 저장하고 다음 분류로 이용하여 분석을 하게 프로그래밍 한다.

4.3 예측 및 분석결과

20개의 분석대상 중 4개의 세그먼트 코드 12, 26, 55, 60은 검색 대상에서 제외되었다. 분석결과Ljung-Box Test를 통한 모형검증은 통과하였으나 MAPE값이 50% 이상의 값으로 부정확한 예측이므로 검색 대상에서 제외되었다.

전체 시장의 규모를 ARIMA(1,1,0)으로 예측한 선행연구[12]와 달리 코드 30, 31, 39, 40, 45, 47, 56은 ARIMA(3,2,1) 모형으로, 코드 46, 53은 ARIMA(3,3,2) 모형으로 예측되었다. 한편 코드 25는 별도의 차분 없이도 분석이 가능한 ARIMA(2,0,2) 모형으로 예측되었다.

표 7. 추정된 모수 및 예측결과
Table 7. The estimated parameters and expected results

코드	모수			2014년 1/4분기 (단위:백만원)		예측율 (%)	2014년 2/4분기 (단위:백만원)		예측율 (%)	MAPE (%)
	p	d	q	실제값	예측값		실제값	예측값		
11	1	3	2	19,454	20,807	106.96	24,339	21,954	90.20	8.38
14	3	3	1	18,167	15,112	83.18	9,848	9,596	97.44	9.69
21	3	2	2	9,127	10,319	113.07	12,152	11,486	94.52	9.27
25	2	0	2	189,709	138,347	72.93	129,500	113,961	88.00	19.54
30	3	2	1	2,169,655	2,317,608	106.82	2,062,990	2,201,254	106.70	6.76
31	3	2	1	18,751	22,999	122.65	28,418	30,238	106.40	14.53
39	3	2	1	132,177	181,271	137.14	193,911	216,798	111.80	24.47
40	3	2	1	360,935	377,440	104.57	358,971	369,968	103.06	3.82
43	2	2	2	246,385	215,919	87.63	218,837	172,289	78.73	16.82
44	3	2	3	44,768	40,199	89.79	19,808	22,273	112.44	11.33
45	3	2	1	38,202	35,078	91.82	22,185	25,311	114.09	11.13
46	3	3	2	40,915	42,398	103.63	58,686	48,544	82.72	10.45
47	3	2	1	76,500	69,021	90.22	55,083	58,169	105.60	7.69
52	0	2	2	32,380	30,631	94.60	30,414	31,142	102.39	3.90
53	3	3	2	17,487	16,839	96.30	47,357	26,032	54.97	24.37
56	3	2	1	145,768	116,599	79.99	84,639	87,049	102.85	11.43

각각의 분류별로 ARIMA 모형의 모수가 표 7과 같이 다르게 나타나기 때문에 G2B 전자상거래 시장규모를 분류별로 예측하려면 전체 시장의 규모의 모수를 추정하기 보다는 해당 부분의 산업분야별로 각각 다른 ARIMA 모형의 모수가 적용되어야 한다. 관측 값과 예측 값의 차이를 비교하는 MAPE 값이 3.90%~24.47%로 나타나 본 연구에서 구현한 ARIMA 모형의 처리방법이 우수한 것을 입증하였다.

V. 결론 및 향후 과제

본 연구는 공공데이터를 이용한 수집, 가공, 분석할 수 있는 방법을 제시하였으며, 대량의 데이터를 가지고 일괄적으로 분석할 수 있는 ARIMA 수요예측모형을 구현하였다.

본 연구와 의의는 다음과 같다. 첫째, 통계청에서 제공된 데이터에 대한 분석과 활용에 그치던 선행연구와 달리 공공데이터 개방 신청을 통하여 연구자가 목적에 맞도록 표 1과 같이 G2B 전자상거래 상품분류의 세분화된 데이터를 수집, 가공 및 분석할 수 있는 방법을 제시하였다.

둘째, 전자상거래 시장규모를 각각의 분류별로 예측하고 모형을 추정하는 선행연구와 달리 오픈소스인 R 프로그램을 이용하여 데이터를 가공하고 일괄적으로 분석할 수 있는 ARIMA 모형을 구현하였다. 또한 대용량의 데이터에서도 처리 할 수 있는 방법을 제안하였다.

셋째, 선행연구에서는 통계청의 데이터에 의존하여 G2B 전자상거래 규모에 대해 포괄적인 수요예측모형을 연구했기 때문에 각각의 분류별 세부시장을 예측하는 데는 한계가 있다. 따라서 본 연구는 각각의 분류를 토대로 보다 세분화된 G2B 산업의 시장규모를 예측하므로, 국내 조달산업의 각 분류별 예측이 가능한 방법을 제시하였다. 또한 국제기준에 맞는 상품분류체계가 적용된 데이터에 대한 연구를 통하여 UNSPSC가 적용된 국가 간의 G2B 전자상거래 시장규모를 비교할 수 있는 효과를 얻을 수 있었다.

본 연구는 공공데이터를 연구자 목적에 맞게 분석하는 방법과 대용량의 데이터를 처리할 수 있는 방법을 제시하였다. 또한 법률에 따라 UN에서 제정한 국제표준 분류코드를 기반으로 하였기 때문에 통계청 제공 데이터로는 분석할 수 없었던 국가 간의 전자상거래 세부시장을 비교할 수 있는 방법을 제안하였다.

본 연구의 한계점으로는 전체 4단계로 구성된 상품분류체계 중 최상위 단계의 분류만 분석하여 실제 하위 단계의 보다 세분화된 분류에 대해 예측모형과는 차이가 있을 수 있는 점이다. 따라서 향후 연구에서는 사용자가 원하는 각 단계별 시계열 모형에 대한 분석이 진행될 수 있으며, ARIMA 모형 이외에 다른 예측모형을 추가하여 적용하는 연구가 필요하다.

References

[1] Public data portal, http://www.data.go.kr

[2] Appleford, Simon, James R. Bottum, and Jason Bennett Thatcher, "Understanding the social web: towards defining an interdisciplinary research agenda for information systems", The DATABASE for Advances in Information Systems, Vol. 45, No. 1, pp. 29-37, Feb. 2014.

[3] Seok-Bong Jeong, "A Novel Approach to Customer Classification and Preference Goods Extraction based on Social Network Analysis for Internet Shopping Malls", Entrue Journal of Information Technology, Vol. 13, No. 1, pp. 57-68, Apr. 2014.

[4] Dae-Gi Kim, Won-Kyun Joo, Eun-jin Kim, and Yong-Ho Lee, "A Case Study on Classification System Design for Public Sector Information Typology", Journal of Digital Policy & Management, Vol. 12, No. 4, pp. 51-68, Apr. 2014.

[5] Box, George EP, and Gwilym M. Jenkins. Time series analysis: forecasting and control, revised ed. Holden-Day, 1976.

[6] Sang-Yun Lee and Hong-Joo Yoon, "The Study on Strategy of National Information for Electronic Government of S. Korea with Public Data analysed by the Application of Scenario Planning",

The Journal of the Korea institute of electronic communication sciences, Vol. 7, No. 6, pp. 1245-1258, Dec. 2012.

[7] Yong-Jun Cho, Yong-Hoon Cho, and Jeong-Uk Kim, "Time Series Analysis of General Marine Fisheries,Journal of Rural Development", Vol. 29, No. 1, pp. 123-134, June 2006.

[8] Keun-Soon Kim and Jae-Hyun Ahn, "A Study on the Real Time Forecasting for Monthly Inflow of Daecheong Dam using Seasonal ARIMA Model", Journal of Korea Water Resources Association, pp. 1395-1399, May 2010.

[9] Pasapitch Chujai, Nittaya Kerdprasop and Kittisak Kerdprasop, "Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models", Proc. IMECS Conf., Hong Kong. Mar. 2013.

[10] Young-Joo Kim, "Study on Low Cost Carrier Demand Forecasting Using Seasonal ARIMA Model", Journal of Tourism Studies, Vol. 26, No. 1, pp. 3-25, Feb. 2014.

[11] Hee-Jae Kim, "The Application of Time Series Analysis under R Environment", Journal of the Korean Data Analysis Society, Vol. 13, No. 1, pp. 331-341, Jan. 2011.

[12] Kyo-Won Choi, "A Study on the Prediction Methods of Domestic e-Commerce Market Size", The Journal of Society for e-Business Studies, Vol. 9, No. 4, pp. 1-17, Nov. 2004.

[13] Man-jai Lee, "Big Data and the Utilization of Public Data", Internet and Information Security, Vol. 2, No. 2, pp. 47-64, Nov. 2011.

[14] The Korea Chamber of Commerce and Industry, http://www.gs1kr.org

[15] Ig-Hoon Lee, "A Study on Supplier Relationship Management System for National Public Procurement", The Journal of Society for e-Business Studies, Vol. 16, No. 1, pp. 101-116, Feb. 2011.

[16] Dong-soo Lee, "Forecasting Quarterly Growth Rates Using IAIP", Journal of The Korean Official Statistics, Vol. 18, No. 2, pp. 66-88, Oct. 2013.

[17] Hyun-ki Park and Jae-kyoung Ahn, "Implementation of the Demand Forecasting Module Linking JAVA and R Programs", Journal of Korean Institute of Information Technology, Vol. 12, No. 3, pp. 85-94, Mar. 2014.

[18] Sang-Kuk Kim, "The Role of Agricultural Cooperative under the Forecast of Korean Apiculture by the ARIMA Model", The Korean Journal of Cooperative Studies, Vol. 25, No. 1, pp. 183-210, Sep. 2007.

[19] Sang-Sok Suh, Jong-Woo Park, Gwang-suk Song, and Seung-Gyun Cho,"A Study of Air Freight Forecasting Using the ARIMA Model", Journal of Distribution Science, Vol. 12, No. 2, pp. 59-71, Feb. 2014.

저자소개

박 현 기 (Hyunki Park)

2005년 2월 : 한국산업기술대학교 컴퓨터공학과(공학사)
2012년 2월 : 서울과학기술대학교 IT정책전문대학원 산업정보시스템(공학석사)
2012년 3월 ~ 현재 : 서울과학기술대학교 IT정책전문대학원 산업정보시스템 박사과정 수료
2005년 ~ 현재 : 새마을금고복지회 전산개발담당
관심분야 : 데이터마이닝, 수요예측, 전자상거래

안 재 경 (Jaekyoung Ahn)

1985년 2월 : 서울대학교 산업공학과(공학사)
1987년 2월 : 서울대학교 산업공학과(공학석사)
1991년 8월 : 아이오와주립대 산업공학과 (Ph. D.)
1991년 10월 ~ 현재 : 서울과학기술대학교 글로벌융합산업공학과 교수
관심분야 : 통신경영, 경제성분석, 기술평가