ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 September: 23(3): 1~27 http://dx.doi.org/10.13088/jiis.2017.23.3.001

뉴스기사를 이용한 소비자의 경기심리지수 생성*

송민채
이화여자대학교 빅데이터분석학
(*nicenara84@naver.com*)

신경식
이화여자대학교 경영대학
(*ksshin@ewha.ac.kr*)

경제주체들의 경기상황에 대한 판단 및 전망은 경기변동에 영향을 미치므로 경기심리지수와 거시경제지표들 간에는 밀접한 관련성을 나타내는 것으로 알려져 있다. 경기선행지표로 국내에서 많이 사용되는 경기심리지수에는 소비자동향조사, 기업경기조사, 경제심리지수가 있다. 그러나 설문조사를 통해 생성된 지수는 자료의 성격상 속보성이 떨어지는 문제가 있다. 본 연구에서는 이러한 정형데이터의 한계를 보완할 수 있도록 비정형데이터에서 정보를 추출해 경기심리지수를 생성하고, 경제분석에서의 활용 가능성을 검토하였다. 민간소비와 관련된 실물지표에는 소매판매업지수와 서비스업생산지수를 사용하였고, 고용지표에는 고용률과 실업률을, 가격지표에는 소비자물가상승률과 가계의 대출금리를 사용하여 지표들 간의 추이 분석 및 시차구조 파악을 위한 교차상관분석을 수행하였다. 마지막으로 이들 지표들에 대한 예측 가능성을 점검하였다. 분석결과, 다른 지표들의 선행지수로 많이 사용되는 소비자심리지수와 비교해 선택 지표들과 높은 상관관계를 보이며, 1~2개월 선행한 것으로 나타났다. 예측력 또한 향상되어 텍스트데이터에서 생성한 소비자 경기심리지수의 유용성이 확인되었다. 온라인에서 생성되는 뉴스기사나 소셜 SNS 등의 텍스트 데이터는 속보성이 뛰어나고, 커버리지가 넓어 특정 경제적 이슈가 발생할 경우 이것이 경제에 미치는 영향을 빠르게 파악할 수 있다는 점에서 경기판단지표로써의 잠재적 가능성이 클 것으로 보인다. 경제분석에서 비정형데이터를 활용한 국내연구는 초기 단계지만 데이터의 유용성이 확인되면 그 활용도가 크게 높아질 것으로 기대한다.

주제어 : 감성 분석, 비정형데이터, 경기심리지수, 소비자동향조사, 경제동향 및 전망

논문접수일 : 2017년 5월 29일 논문수정일 : 2017년 7월 31일 게재확정일 : 2017년 9월 13일
원고유형 : 일반논문 교신저자 : 신경식

1. 개요

가계와 기업의 경기에 대한 판단 및 전망은 생산, 소비, 투자의 변화를 통해 단기적 경기변동을 발생시키므로 경제주체들이 느끼는 체감경기와 거시경제지표들 간에 밀접한 관련성을 보이는 것으로 알려져 있다(Bram and Ludvigson (1997); Ludvigson (2004); Gelpera et al (2007)). 국내의 대표적인 경기심리지수에는 한국은행이 매달 설문조사를 통해 경제주체들의 경기인식과 전망을 파악하여 발표하는 소비자동향조사(Consumer Sentiment Index), 기업경기조사(Business Survey Index), 경제심리지수(Economy Sentiment Index)가 있다. 그러나 이러한 지수들은 설문조사 항목에 관한 정보만 수집되어 지수에 반영되므로 특정 이슈가 발생할 경우 그것이

* 이 연구는 2014학년도 이화여자대학교 교내연구비 지원에 의한 연구임 (This work was supported by the Ewha Womans University Research Grant of 2014).

경기인식에 미치는 영향을 알 수 없다는 문제가 있다. 또한 조사 및 수집, 집계까지 상당한 시간이 소요되어 매월 말 발표되므로 자료의 이용 가능 시점에 제약이 있다. 본 연구에서는 이러한 정형데이터의 한계를 보완할 수 있도록 뉴스기사에서 정보를 추출하여 소비자의 경기심리지수를 생성한 후 그 유용성 확인을 목적으로 한다. 표본조사 및 수집 등을 통해 만들어지는 수치 중심의 정형 데이터와 달리 구조화되지 않은 비정형데이터는 데이터의 수집 단계부터 처리 및 분석방법까지 매우 상이하다. 따라서 비정형데이터의 특성을 고려한 연구설계가 중요하다.

비정형데이터를 활용한 다양한 연구방법 중 본 연구에서 감성분석을 선택한 것은 감성분석의 분류 문제(Classification problem)가 소비자동향조사에서 경기심리지수를 계산하는 방법과 유사하기 때문이다. 소비자심리지수는 경기에 대한 가계의 인식을 긍정, 부정, 중립으로 조사하여 이를 지수화한 지표이다. 이와 유사하게 감성분석 역시 텍스트에 드러난 감성을 긍정과 부정, 중립으로 분류한다.

한편 텍스트에 대해 감성분석을 수행한 기존 연구들은 온라인 상품평이나 트위터, 블로그와 같은 자료를 대상으로 한 경우가 많았다. 이는 주관적 문장으로 구성된 텍스트들이 감성을 내포할 가능성이 높다고 보기 때문이다(Balahur et al, 2010). 그에 반해 뉴스기사를 이용한 감성분석에 관한 연구는 많지 않은데 이는 객관적 입장에서 작성된 텍스트는 감성전달이 불가능할 것으로 보아 연구대상에서 제외된 측면도 크다. 그러나 이러한 텍스트 역시 감성분석이 가능하다. 뉴스기사가 자신의 경험이 아닌 제 삼자의 입장에서 사건이나 상황을 기술하므로 주관적 텍스트에 비해 감성어휘의 등장 빈도가 낮을 수 있으나 감성을 표현하는 어휘들이 사용된다. 또한 인터뷰나 인용 등을 통해 특정인의 의견이나 감성이 전달될 수 있다(B. Liu, 2015).

이와 별개로 한 문서 내에서 둘 이상의 개체(Entity) 또는 속성(Aspect)에 대해 상이한 감성이 표현되거나 다수의 주체가 감성을 표현하는데 그 극성이 다를 경우 이를 구분하지 않으면 감성분류의 정확성이 떨어지는 문제가 발생한다(Aggarwal and Zhai, 2012). 소셜 SNS나 상품평과 같은 텍스트는 감성을 표출하는 주체가 1인이고, 주로 한 속성에 대해서만 언급하므로 이로 인한 오류가 상대적으로 적을 수 있다. 이러한 점도 연구 분석대상의 선정에 영향을 주었다. 반면, 뉴스기사는 객관적 입장에서 작성되는 텍스트이어야 하므로 한 기사 내에서도 상반된 입장을 같이 전달하는 경우가 많다. 본 연구는 한 기사의 전반적 감성을 분류하는 문서 단위(Document level)의 감성분석이 아닌, 기사에 등장한 감성어휘의 빈도 수를 이용해 가계의 경기심리를 판단하므로 이러한 문제로 발생하는 오류는 크지 않을 것으로 판단된다.

이렇게 비정형데이터에서 추출된 정보로 만들어진 소비자의 경기심리지수가 실제 활용될 수 있기 위해서는 다른 경제지표와의 연관성이 매우 중요하다. 본 연구에서 생성한 경기심리지수의 유용성을 점검하기 위해서 가계 부문의 경제활동과 밀접하게 관련된 실물경제지표와 고용지표, 가격지표를 선택하여 경기심리지수와 지표들 간 추이분석 및 교차상관분석을 수행하였다. 민간소비와 관련된 변수에는 소매판매업지수와 서비스업생산지수를 사용하였고, 고용활동은 고용률과 실업률, 가격변수에는 소비자물가상승률과 가계의 대출금리를 적용하였다. 마지막으로 이들 지표에 대한 예측력을 확인하기 위해 단순

회귀분석을 하였다. 그 결과, 뉴스기사에서 생성된 소비자의 경기심리지수가 벤치마크지수인 소비자심리지수에 비해 높은 상관계수를 보이며, 1~2개월 선행한 것으로 나타났다. 또한 소비자심리지수와 비교해 다른 경제지표에 대한 예측력이 향상되어 경기심리지수로써 유용함이 확인되었다.

온라인에서 생성되는 뉴스기사나 소셜 SNS 등의 텍스트 데이터는 속보성이 뛰어나고, 커버리지가 넓어 경제 관련 이슈가 발생할 경우 이것이 경제에 미치는 영향을 빠르게 파악할 수 있다는 점에서 경기판단지표로써의 잠재적 가능성이 클 것으로 보인다. 경제분석에서 비정형데이터를 이용한 연구는 많지 않지만 데이터의 유용성이 보장되면 그 활용도는 크게 높아질 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2장에서는 비지도학습을 통한 감성분석 연구와 텍스트를 경제분석에 적용한 선행연구를 중심으로 살펴본다. 3장에서는 뉴스기사의 수집, 데이터의 전처리 과정 및 감성사전 구축방법, 소비자의 경기심리지수 생성방법, 비교 분석에서 사용할 주요 경제지표에 대해 설명한다. 4장에서는 본 연구에서 생성한 소비자의 경기심리지수에 대한 유용성을 점검한 분석 결과를 다룬다. 마지막으로 5장에서 연구 내용을 요약하고, 향후 연구과제를 제시하며 마무리한다.

2. 관련 연구

감성분석(Sentiment analysis)이란 텍스트에 표현된 개체와 그 속성에 대한 의견, 감성, 평가, 태도 등을 분석해 텍스트에 드러난 감성을 분류하는 것이다(Pang et al, 2002). 이러한 감성분석은 분석 데이터에 레이블(Label)이 있는 경우와 그렇지 않는 경우에 따라 크게 지도학습(Supervised learning)과 비지도학습(Unsupervised learning)으로 나눌 수 있다. 지도학습은 도메인에 적합하게 데이터를 학습시켜 자동으로 분류문제를 수행하기 때문에 비지도학습에 비해 성과가 높은 것으로 나타났다(Pang and Lee, 2008). 따라서 온라인 상품평이나 영화 평점처럼 별점이나 만족도 점수 등을 통해 텍스트의 감성이 이미 분류된 레이블 있는 데이터라면 우선적으로 지도학습을 고려해볼 수 있다. 그러나 레이블이 부착된 대량의 데이터를 현실에서 찾기란 매우 어렵다. 본 연구의 분석대상인 뉴스기사 역시 레이블이 없는 데이터로, 이렇게 레이블이 없는 데이터는 비지도학습만 가능하다. 비지도학습을 통한 텍스트의 감성분석에는 단어 간 상관관계 분석을 통해 문장의 극성을 구분하는 방법, 자연어처리 방식이나 문장의 패턴을 이용하는 방법, 극성이 부여된 단어들로 구성된 감성사전을 이용해 분류하는 방법 등이 있다. 본 연구에서는 비지도학습에서 많이 사용되는 감성사전을 활용하여 감성을 분류하였다.

2.1 비지도학습을 통한 감성분석

비지도학습을 통한 감성분석에는 크게 구문패턴(Syntactic pattern)의 활용과 감성어휘를 이용하는 접근방법(Sentiment lexicon-based approach)이 있다. 전자의 대표적인 초기 선행연구에는 Turney(2002), Turney and Littman(2003)이 있다. 이 연구들은 통계적 기법인 PMI(Point wise mutual information)를 이용해 텍스트의 감성을 분류했다. PMI는 문서 내 특정 단어들이 동시

에 등장할 확률을 계산하는 방법으로, 극성이 유사할 경우 PMI 또한 높은 값을 가질 것이라는 가정 하에 두 단어 간의 유사성을 추정하는 것이다.

또 다른 비지도학습은 감성사전을 활용해 감성을 분류하는 것이다. 일반사전과 달리 감성사전은 감성을 전달할 수 있는 어휘들로 구성되며, 어휘의 극성이 주된 정보이기 때문에 감성사전에 등록된 어휘들의 감성을 먼저 분류해야 한다. 어휘의 감성을 어떻게 분류하느냐에 따라 감성사전을 구축하는 방법도 다양하다. 사전기반(Dictionary based) 방식은 WordNet과 같이 일반사전에 등록된 형용사의 유의어와 반의어 정보를 통해 사전을 구축한다(Kamps et al(2004); Hu and Liu(2004)). 이들은 특정 형용사의 유의어는 동일한 감성을, 반의어는 반대의 감성을 공유한다는 것을 가정하고 있으며, 적은 수의 어휘에 대해서만 감성을 분류해 놓으면 이를 기반으로 비교적 쉽게 감성사전을 만들 수 있다는 장점이 있다. 그러나 도메인이나 문맥에 따라 감성의 극성이 바뀌는 중의적 어휘는 고려하지 못하기 때문에 분류의 정확성이 떨어지는 문제가 발생한다(Hoang et al(2008); Neviarouskaya et al(2009); Mohammad et al(2009)). 선행연구에서도 범용의 감성사전(General sentiment lexicon dictionaries)으로 분석을 수행하는 것보다 도메인에 적합하게 극성이 분류된 감성사전(Corpus based approach)을 사용할 경우 분석성과가 개선된 것으로 나타났다(Huang et al, 2014).

이후 범용의 감성사전 문제를 해결할 수 있는 다양한 방법들이 제안되었다. Wilson et al(2005)은 2단계 지도학습기법을 적용하여 문맥에 따라 극성이 바뀌는 특징 어휘들(Feature set)을 고려했고, Kanayama and Nasukawa(2006)은 문맥 내(inter)와 문맥 외(intra)에서 동시 등장하는 감성어휘를 구분해 감성사전을 구축했다. Kennedy and Inkpen(2006), Taboada et al(2011)은 문맥에 따라 문장의 감성을 바뀌게 하는 'not good'과 같은 negation 어휘들을 별도로 처리하여 감성분석을 수행했다. Lu et al(2011)은 속성에 따라 감성어휘의 극성이 바뀌는 문제를 고려한 감성사전(Context dependent sentiment lexicon)을 구축하여 분석하였다. 그 결과 공통적으로 범용의 감성사전이 가지는 문제를 고려할 경우 분류의 예측력이 높아지는 것으로 나타났다.

2.2 비정형데이터를 활용한 경제분석

비정형데이터를 경제분석에 적용한 선행연구는 아직 많지 않은데, 최근 들어 비정형데이터의 관심이 커지면서 텍스트 데이터를 이용한 연구들이 나오고 있다. Lee and Hwang(2014)에서는 검색을 일종의 자발적인 조사에 대한 응답으로 보고, 소비자심리지수와 유사한 방식으로 네이버 검색 경기지수를 생성하여 다른 경제지표와의 비교 분석을 수행하였다. 경기지수 생성에는 네이버가 제공하는 검색 포털에서 '경기 호황'과 '경기 불황'으로 검색하여 추출된 검색 통계를 사용하였다. 네이버 검색 경기지수의 유용성을 확인한 결과 다른 심리지수와 유사하게 움직이며, 경기동행지수 순환 변동치와 밀접하면서 선행지수인 것으로 나타났다. 또한 경제성장률과 민간소비 증가율에 대한 예측력을 확인한 결과, 특히 금융위기가 발생한 기간에 우수한 성과를 보였다. Hwang(2015)은 웹소셜 플랫폼에서 제공하는 SNS 데이터를 이용해 경제상황에 대한 소비자의 경기인식을 측정했다. 이 연구에서는 경제 상황과 관련된 감성사전을 별도로 구축한 후

감성사전에 기초하여 소비자들의 경제상황에 대한 긍정 또는 부정 어휘의 빈도 수를 산출하여 경기심리지수를 생성했다. 그 결과 소비자심리지수, 경제심리지수, 경기동행지수 등과 밀접한 관련성을 보인 것으로 나타났다. 또한 교차상관분석에서는 소비자심리지수와 동행하면서 높은 상관관계를, 경기동행지수와는 9개월 후행하면서 높은 상관관계가 나타났다. 이 외에 텍스트를 경제분석에 적용하여 주가예측을 시도한 연구들(Bollen et al(2011); Li et al(2014); Nguyen et al(2015))이 있으나 뉴스기사에서 경기심리를 판단하여 이를 실물경제지표와 분석한 연구는 찾아보기 어렵다. 경기판단지표로써 텍스트 데이터의 유용성이 확인되면 경제분석에서 다양하게 활용될 수 있을 것으로 기대한다.

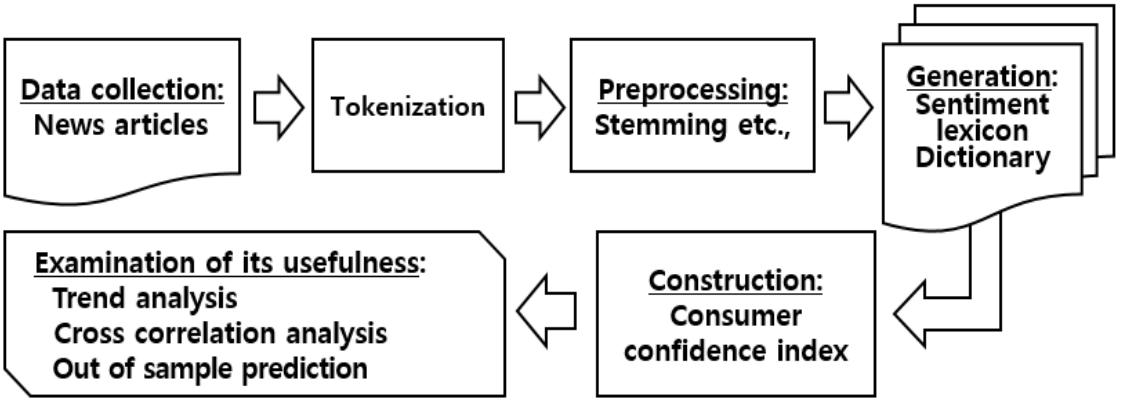
3. 연구방안

3.1 연구절차

본 연구의 분석단계는 <Figure 1>과 같다. 먼저 뉴스기사를 수집하고, 수집된 데이터는 전처리 과정을 거친다. 전처리 후 생성된 데이터는 감성사전에 등록할 후보 군과 지수 생성에 사용된다. 감성사전이 구축되면 전처리 과정을 거친 데이터에 적용한다. 감성사전에 등록된 어휘일 경우 지수 계산에 사용되며, 감성어휘가 아닐 경우 쓰이지 않는다. 이를 토대로 월별 소비자의 경기심리지수를 만들고, 새로 생성된 지수는 가계의 경제활동과 관련된 지표들과의 비교 분석을 통해 그 유용성을 점검한다.

3.2 데이터 수집

데이터 수집에는 상당한 시간이 소요되므로 언론사는 국내 점유율 1위 포털 사이트인 네이버 뉴스에서 제공하는 11개 종합일간지와 9개 경제일간지로 한정하였다. 본 연구의 분석기간은 2015년 1월부터 2017년 3월이다(<Table 1>). 뉴스데이터는 자료 수집이 수월하고, 비속어나 신조어가 적고, 문법을 준수하려는 경향이 있어 전처리가 비교적 단순하다는 장점이 있다. 키워드는 본 연구의 벤치마크 지수인 소비자심리지



〈Figure 1〉 Proposed method of flow chart

〈Table 1〉 Data collection: 2015.1~2017.3

Keyword	Households business	Households income	Households expenditure	Total
Number of articles	107	312	173	592

수의 조사항목[1]을 반영하여 '가계 경기'와 '가계 소득', '가계 소비'로 선정하였다. 검색은 기사의 본문이 아닌 기사제목에서 키워드가 등장한 기사를 수집하였다. 이는 본문으로 분석대상을 설정하면 크게 두 가지 문제가 발생할 수 있다고 판단되었기 때문이다. 먼저 한 기사 내에서 다양한 사안을 다루는 기사들이 많고, 검색 키워드가 특정 주제에 대해 한정하여 사용되는 용어라기보다는 일반적으로 사용되는 단어이므로 해당 키워드가 본문에서 등장한 기사로 설정할 경우 가계의 경기인식과 무관한 내용들까지 분석대상에 들어오는 문제가 심각해진다. 다음은 동일 이슈 또는 사건에 대해 언론사마다 비슷한 내용을 보도하는 경향이 있어 가급적 중복된 내용의 기사는 피하기 위해서이다. 검색 키워드가 기사제목에 등장한 경우로 분석대상을 좁히면 이러한 문제가 크게 줄어듦을 확인하였다.

3.3 전처리(Preprocessing)

언어는 그 성격 상 시간이 지남에 따라 변하므로 자연어라 한다. 자연어 처리는 대상 언어에 따라 분석 방식이 상당히 달라진다. 수치 중심인 정형 데이터는 분석에 용이하게 구조화되어 있지만 텍스트는 그렇지 않기 때문에 전처리 단계가 필요하다. 전처리를 거친 데이터가 기초자료로 사용되기 때문에 분석에서 매우 중요한 단계이다.

본 연구에서는 띄어쓰기가 의미를 가지는 최소 단위로 보아 띄어쓰기를 기준으로 어절 분리(Tokenization)를 수행하였다[2]. 일반적으로 불용어(Stop word)는 감성의 분류에 영향을 미치지 않다고 보기 때문에 제거하는데, 본 연구에서 이러한 단어는 감성사전에 등록하지 않아 자동으로 삭제되므로 별도의 불용어 처리(Stop word removal)는 하지 않았다. 어간 처리(Stemming) 단계에서는 한글의 특성을 고려해 최소한의 어간 처리만 하였다. 언어유형학(Linguistic typology)의 분류에 따르면 한글은 교착어, 영어는 굴절어에 해당한다. 굴절어는 어간(Stem)과 어형을 변형시키는 어미(Suffix)로 구성되는 반면, 교착어는 하나의 어근(root) 혹은 어간에 각각 단일한 기능을 가지는 하나 이상의 접사(Affix)로 이루어져 있다. 굴절어의 경우 어근에 붙을 수 있는 어미 유형이 한정되어 있어 몇 가지 유형만 알면 영어의 어간 처리는 비교적 단순하여 자동화가 용이하다. 그러나 한글은 그 특성 상 한 어근에 여러 접사가 붙을 수 있기 때문에 어간 처리가 매우 까다롭다. 어간 처리 과정에서 발생할 수 있는 오류를 줄이기 위해 본 연구에서는 '을',

1) 세부적으로 소비자심리지수는 현재 생활형편지수, 생활형편전망지수, 가계수입전망지수, 소비지출전망지수, 현재경기판단지수, 향후 경기전망지수로 구성된다. 이때의 현재는 6개월 전과 비교하여, 전망은 현재와 비교한 6개월 후를 기준으로 조사대상인 가구주의 생활형편, 가계수입, 소비지출, 경기판단에 대한 인식이다.

2) 한글은 어절과 단어의 형태가 일치하는 경우가 많지 않기 때문에 어절과 단어를 동일한 개념으로 보기 어렵다. 그러나 영어의 단어(word)와 가장 가까운 개념이 한국어의 어절에 해당한다.

‘를’과 같이 문장에서 공통적으로 많이 사용되는 소수의 접사만 제거하였다. 마지막으로 특수문자, 영어, 숫자, 1 글자의 단어는 제외하였다. 전처리 과정을 거치고 중복된 어휘들을 제거하여 감성사전에 등록될 어휘의 후보 군과 지수 계산에 사용할 기초 데이터를 생성시켰다[3].

3.4 감성사전 구축

감성사전은 감성분석의 성과에 높은 영향을 미치는 중요한 요소 중 하나이다. 감성사전의 구축에는 문장 또는 단어의 극성이 미리 정의된 감성사전을 바탕으로 새로운 어휘들의 유사성 및 거리 관계 등을 통해 어휘를 추가해 가는 방식과 실제로 수집된 문장들에 대한 형태소 분석을 바탕으로 개별 단어의 극성을 정의하여 구축하는 방법 등이 있다. 영어의 경우 일반사전인 ‘WordNet’에 감성 정보를 부착한 ‘SentiWordNet’과 같은 범용적 감성사전이 선행연구에서 많이 사용된다. 그러나 한글은 공개적으로 많이 사용되는 감성사전이 없어 본 연구에서는 별도로 구축하였다.

일반적인 사전이 단어의 정의 및 품사 등이 중심이라면, 감성사전은 어휘의 극성 값이 주된 정보가 된다. 본 연구에서는 긍정, 부정, 중립으로 극성을 분류하였으며, 어휘는 Unigram 기반으로 하였다. Unigram 기반은 단순히 단어주머니(Bag of words)만 이용하므로 문장 구조에 내포된 감성 정보는 반영하기 어렵다. 또한 ‘매우 훌륭한’과 같이 단어의 연결을 통해 감성의 강도가 표현되는 어휘들은 별도의 처리하지 않으면 이를 고려하지 못한다는 한계가 있다. 한편 감성사전 생성에는 많은 시간과 노력이 들지만 본 연구처럼 해당 도메인에 적합하게 감성을 분류할 경우 중의적 어휘나 도메인에 따라 극성이 달라지는 도메인 의존성을 가지는 감성어휘로 인한 오류의 발생 가능성은 낮아진다. 본 연구에서 구축한 감성사전의 구성은 <Table 2>와 같다[5].

3.5 뉴스기사를 통한 소비자의 경기심리지수 생성

본 연구에서는 소비자동향조사와 유사한 방식으로 경기심리지수를 생성하였다. 구체적으로는

〈Table 2〉 Construction of Sentiment lexicon dictionary

Polarity	Number of words	Importance	Composition of Part of speech[4]
Negative	5,385	22% (64%)	N=3,438 (64%), P=781 (15%), M=767 (14%), I=283 (5%)
Positive	2,981	12% (36%)	N=1,599 (54%), P=540 (18%), M=555 (19%), I=214 (7%)
Neutral	16,677	67% (-)	N=11,447 (69%), P=1,713 (10%), M=2,029 (12%), I=1,106 (7%)
Total	25,043	100% (-)	N=Noun, P= Verb and Adjective, M= Adverb, I= Interjection

3) R에서 KoNLP 패키지 내 extractNoun 코드를 적용할 경우 명사 위주로 추출되기 때문에 본 연구에서는 그 코드를 사용하지 않고, R의 strsplit 코드로 이용해 어절을 분리하였다. 이후 R의 KoNLP 패키지를 사용하여 전처리하여 분석 데이터를 생성하였다.

4) R에서 KoNLP 패키지 내 Simplepos09 코드를 사용하여 어휘의 품사를 확인하였다.

5) 전처리 후 생성된 어휘들에 대해 경제학을 전공하고 경제 분야 국책연구소에서 5년 이상 근무하고 있는 연구원 3명이 어휘의 극성을 분류하여 감성사전을 구축하였다.

$$\text{소비자의 경기심리지수(NSI)} = \frac{\sum(\text{긍정어휘수}\times 1+\text{중립어휘수}\times 0+\text{부정어휘수}\times(-1))}{\sum(\text{긍정 어휘 수}+\text{중립 어휘 수}+\text{부정 어휘 수})}\times 100+100$$

뉴스기사에서 추출된 어휘 중 감성사전에 등록되어 있는 긍정어휘와 부정어휘, 중립어휘의 개수를 취합한 후 긍정어휘에 대해서는 1를, 중립어휘는 0, 부정어휘에 대해서는 -1의 값을 부여하였다. 다른 경제지표와의 비교를 위해 본 연구에서는 월별 기준으로 생성하였으나 일별, 주별 등 기사의 수집 빈도에 따라 다양하게 확장할 수 있다. 분석기간 동안 수집된 월별 기사의 수와 기사별 등장하는 어휘의 수가 편차가 있으나 지수 계산 과정에 이러한 차이가 조정되기 때문에 지수에 대해 별도로 정규화(Normalization) 조정은 하지 않았다.

3.6 주요 경제지표

3.6.1 실물지표: 소매판매업지수, 서비스업생산지수

실물경제에 관한 가장 공신력 높은 통계는 국내총생산(GDP)이다. GDP의 50%를 차지하는 민간소비는 그 형태에 따라 소비재인 내구재, 준내구재, 비내구재와 서비스로 구성된다(<Table 3>). 그러나 GDP 잠정치는 매분기 종료 후 3개월 이내 분기별로 발표되어 경기동향을 판단할 수 있는 지표로써 속보성이 크게 떨어진다. 이러한 문제로 국내통계를 조사·집계하는 통계청은 '산업활동조사'를 통해 매월 말 여러 경기 관련 통계를 발표하고 있다. 산업활동조사 항목에서 민간소비와 밀접한 관련성을 보이는 지표는 소매판매업지수와 서비스업 생산지수이다.

소매판매액통계는 가계의 소비동향을 파악하기 위해 작성되는 지표로써 상품의 사용기간과 가격의 고저에 따라 내구재, 준내구재, 비내구재로 분류한다[7]. 조사대상은 개인 및 소비용 상품을 일반대중에게 판매하는 약 2,700개 표본 기업체이다. 소매판매업지수는 월별 소매 상품판매액을 2010년 월 평균 상품판매액으로 나눈 가중치를 적용하여 작성된다. 서비스업생산지수는

〈Table 3〉 Components of private consumption (Fiscal year 2016, Wbn)

GDP	Consumption	Private consumption[6]	Durables	Quasi durables	Nondurables	Services
1,637,421	1,047,483	798,364	73,780	63,435	188,940	421,749
100.0%	64.0%	48.8%	4.5%	3.9%	11.5%	25.8%
	100.0%	76.2%	7.0%	6.1%	18.0%	40.3%
		100.0%	9.2%	7.9%	23.7%	52.8%

Source: Bank of Korea

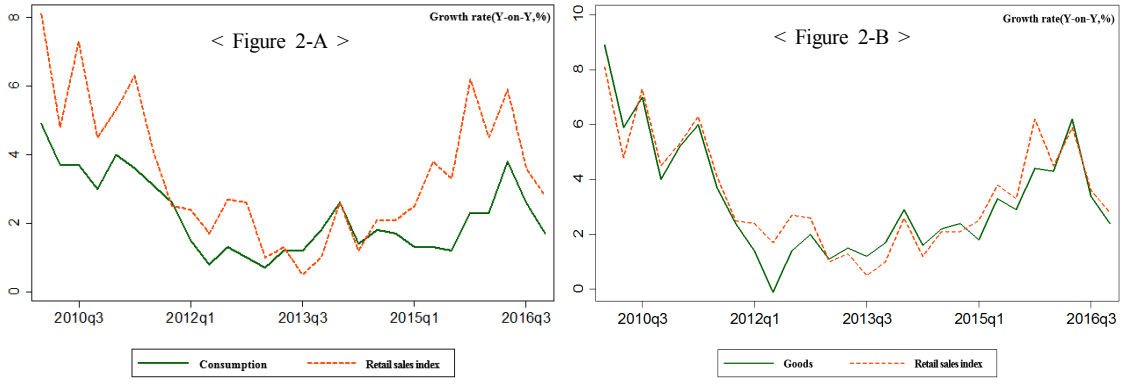
6) 국외소비지출 등이 있어 소비재와 서비스의 합계가 민간소비지출과 일치하지 않는다.

7) 내구재란 1년 이상 사용이 가능하고 주로 고가인 승용차, 가전제품, 컴퓨터 및 통신기기, 가구 등의 상품으로 구성되며, 준내구재는 1년 이상 사용이 가능하나 주로 저가인 의복, 신발 및 가방, 운동 및 오락용품 등의 상품, 비내구재는 주로 1년 미만 사용되는 음식료품, 의약품, 화장품, 서적 및 문구, 차량연료 등의 상품이다.

서비스업 전체[8] 및 개별 업종의 생산활동 파악을 목적으로 작성된다. 개별 업종의 상대적 중요도를 고려해 월별 매출액에 대해 2010년 월 평균 매출액으로 나눈 가중치를 적용하여 지수화한 자료이다. 동 지수는 약 10,200개 표본 기업체를 대상으로 한다.

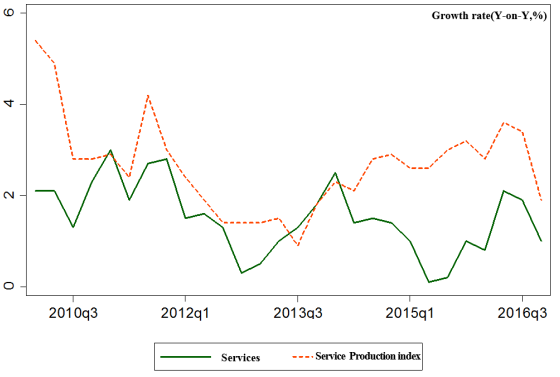
이러한 지수는 지수 조정방법에 따라 경상지수와 불변지수, 계절조정지수가 있다. 경상지수는 물가를 반영하지 않는 지수이고, 불변지수는 경상지수를 물가지수로 나눈 값이다. 계절조정지수는 불변지수에 대해 X-13ARIMA-SEATS 방식을 적용해 계절조정인자를 제거해 생성된다. 일반적으로 경기동향 및 단기적인 경기변동 파악을 목적으로 하는 경우에는 계절조정지수를, 성장수준 분석에는 원지수가 많이 사용되므로 본 연구에서는 계절조정지수를 선택하였다. 경기심리지수와의 비교 분석에서는 두 지수 모두 전년동월대비 증가율로 변환한 값을 사용하였다.

GDP 내 민간소비와 소매판매업지수, 서비스업생산지수 간 관련성을 보기 위해 먼저 추이를 확인하였다. GDP는 분기별 자료만 이용 가능하므로 분기 기준으로 비교하였다. <Figure 2-A>는 민간소비와 소매판매업지수의 증가율 추이이다. 민간소비에는 소비재 외에 서비스와 가계의 국외소비지출 등도 포함되므로 두 시계열의 방향성은 일치하나 증가율의 수준에는 차이가 있다. 다음으로 소비재만 고려할 경우 소비재와 소매판매업지수 간 증가율 추이는 추세뿐 아니라 수준도 매우 유사해 가계의 소비재 지출 동향을 판단할 수 있는 지표로 소매판매업지수가 유용함을 확인할 수 있다(<Figure 2-B>). 한편 민간소비의 50%를 차지하는 서비스와 서비스업 생산지수의 추이를 비교해 보면, 추세는 전반적으로 일치하지만 특정 기간(2014년 3분기~2016년 3분기)에 증가율의 수준에 괴리가 크게 발생하였다(<Figure 3>). 이 시기와 본 연구의 분석기간(2015년 1월~2017년 3월)이 일정 기간 겹치므로 서비스업생산지수와 새로 생성한 지수 간의 비교 시 이 점을 유념할 필요가 있다.



〈Figure 2〉 Consumption Expenditure and Goods compared to Retail sales index of trends

8) 모집단은 2010년 기준 경제총조사 중 13개 서비스업 대분류(제9차 한국표준산업분류 기준)에 해당되는 기업체이다.



〈Figure 3〉 Service compared to Service production index of trends

3.6.2 고용지표: 고용률 및 실업률

고용지표는 경제의 수요와 공급 측면을 모두 반영하므로 경제동향 파악에 매우 중요한 정보를 제공한다. 통계청은 만 15세 이상인 자의 경제활동상태를 파악하여 매월 말 '고용동향'을 발표하고 있다. 고용동향의 주요 항목은 경제활동참가율과 고용률, 실업률이다. 본 연구에서는 가계의 고용활동을 반영하는 지표로 고용률과 실업률을 사용하였다. 고용률은 15세 이상 인구에서 취업자가 차지하는 비중으로 구성되며, 실업률은 취업자와 실업자의 합계인 경제활동인구 중 실업자가 차지하는 비율이다[9]. 4장의 비교 분석에서는 고용률과 실업률의 전년동월 대비 증감 값을 사용하였다.

3.6.3 가격지표: 소비자물가상승률 및 가계의 대출금리

물가란 시장에서 거래되는 개별 상품들의 가격 및 서비스의 요금을 경제생활에서 차지하는 중요도를 고려하여 평균한 종합적인 가격수준을 말한다. 소비자물가지수는 가계가 주로 구입하는 상품과 서비스의 가격 수준을 측정하는 지수로써 언론에서 자주 접하는 인플레이션률에 해당한다. 본 연구에서는 소비자물가지수의 전년동월 대비 증가율을 사용하였다.

한편 가계의 경기판단에 많은 영향을 줄 것으로 예상되는 또 다른 가격지표로 가계의 대출금리를 선택하였다. 신규취급액 기준의 금리가 최근 동향 파악에 사용되므로 본 연구에서도 신규취급액 기준을 적용하였다.

3.6.4 가계의 경기심리지표: 소비자심리지수

가계를 대상으로 한 심리지수에는 한국은행이 조사·발표하는 소비자동향조사가 있다[10]. 소비자심리지수는 소비자들의 경제상황에 대한 심리를 종합적으로 나타내는 지표로써 소비자동향조사 항목 중 경제지표와의 상관성 및 선행성이 높은 6개 주요 개별지수(각주 1 참고)를 이용해 계산된다. 동 지수는 장기평균치(2003년 1월~2016년 12월)를 기준으로 하여 기준 값인 100 보다 크면 낙관적임을, 100 보다 작으면 비관적으로 해석한

9) 경제활동인구는 만 15세 이상 인구 중 조사대상 기간 동안 상품이나 서비스를 생산하기 위하여 실제로 수입이 있는 일을 한 '취업자'와 일을 하지는 않았으나 구직활동을 한 '실업자'로 구성된다. 취업자는 조사대상 주간에 수입을 목적으로 1시간 이상 일한 자 등으로 구성되며, 실업자는 조사대상 주간에 수입 있는 일을 하지 않았고, 지난 4주간 일자리를 찾아 적극적으로 구직활동을 하였던 사람으로 일자리가 주어지면 즉시 취업이 가능한 자로 정의된다. 비경제활동인구는 만 15세 이상 인구 중 조사대상 기간에 취업도 실업도 아닌 상태에 있는 자이다.

10) 소비자동향조사는 매월 15일 전후 1주일 간 전국 도시 2,200 가구주를 대상으로 가계의 경제상황에 대한 인식과 향후 소비지출전망 등을 설문 조사하여 그 결과를 지수화한 자료이다. 설문조사는 응답자의 심리강도가 반영되도록 5점 척도(매우 긍정, 다소 긍정, 비슷, 다소 부정, 매우 부정)로 구성되며, 그에 따라 각각 다른 가중치(1, 0.5, 0, -0.5, 1)를 부여한다.

$$\text{소비자심리지수(CSI)} = \frac{\sum((\text{매우긍정}\times 1+\text{다소긍정}\times 0.5+\text{비슷}\times 0)-(\text{다소부정}\times 0.5+\text{매우부정}\times 1))}{\text{전체 응답 가구수}}\times 100+100$$

다. 본 연구에서는 소비자심리지수를 벤치마크 지수로 선택하였다.

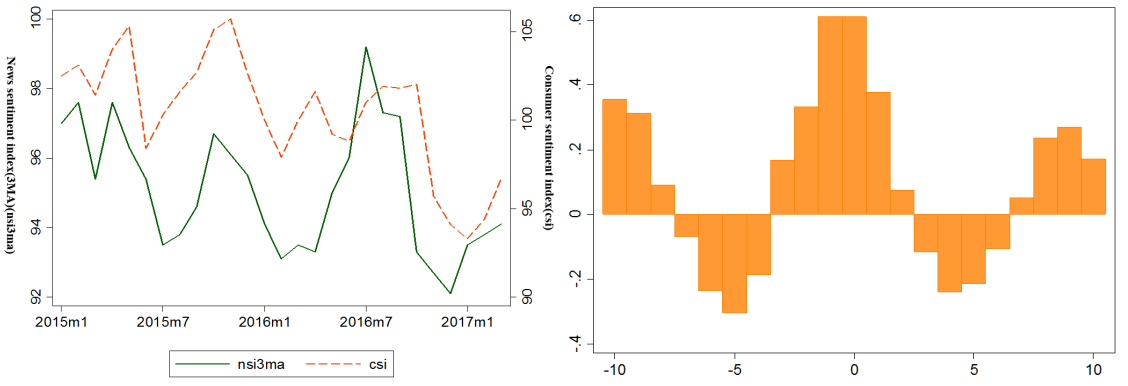
4. 분석결과

본 연구에서는 뉴스기사를 통해 생성한 소비자의 경기심리지수가 유용한지 확인하기 위해 추이 및 교차상관분석을 실시하고, 주요 경제지표들에 대한 예측 가능성을 점검하였다. 교차상관분석에서는 교차상관계수의 최대값을 기준으로 시차구조를 시각화하였다. 예측력 분석에는 소비자심리지수를 예측변수로 사용했을 때와 비교해 모형의 예측력 및 적합도 향상 여부를 검토하였다.

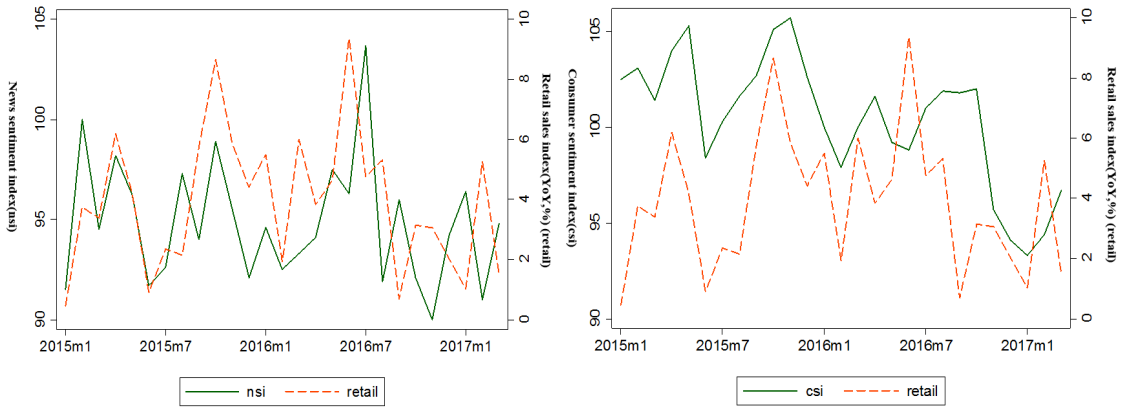
4.1 추이분석

가계의 경제활동에 관한 경제지표와의 관계를 살펴보기에 앞서 정형데이터로 작성된 소비자심리지수(CSI)와 뉴스기사인 비정형데이터에서 추출된 정보로 생성된 경기심리지수(NSI) 간의 관계를 분석하였다. NSI는 CSI에 비해 변동성이 크게 나타나 NSI를 3개월 평균한 후 두 지수 간의 추이를 비교하였다(<Figure 4>). 그 결과 두 지수 간 높은 관련성을 확인할 수 있었다. 한편 3개월 평균법을 적용할 경우 상대적으로 변동성이 작아지는 효과도 있지만 NSI가 CSI의 선행지표라는 해석도 가능하다. CSI의 t 시점 값에 대해 NSI의 t, t-1, t-2기 값 정보가 대응되기 때문이다. 이를 확인하기 위해 두 지수 간에 교차상관분석을 수행하였다. <Figure 4>에서 보듯 NSI가 CSI보다 2개월 선행하며 양(+)의 방향으로 높은 상관관계를 보였다.

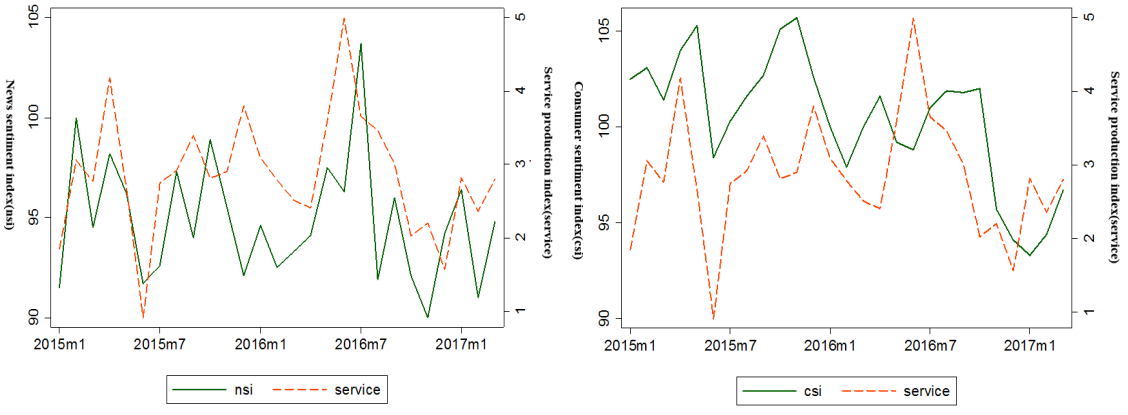
다음으로 가계의 경기심리지수와 소매판매업지수, 서비스업생산지수 간의 관계를 살펴보았다. <Figure 5>는 소매판매업지수와, <Figure 6>는 서비스업생산지수와의 추이를 비교한 것이다. 그림에서도 확인할 수 있듯이 CSI와 비교해 NSI는 두 지수와 보다 밀접하게 움직이는 것으



〈Figure 4〉 NSI (3MA) vs. CSI: Trends and Cross correlation analysis



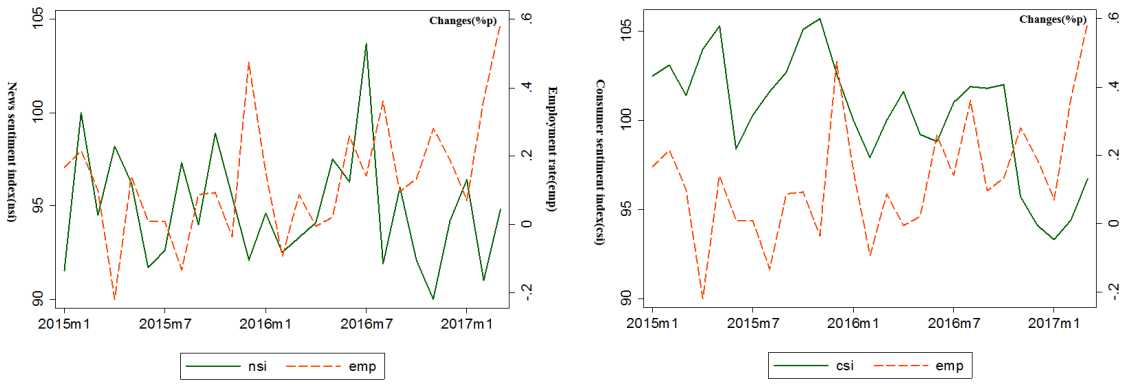
〈Figure 5〉 Retail sales index: NSI vs. CSI of trends



〈Figure 6〉 Service production index: NSI vs. CSI of trends

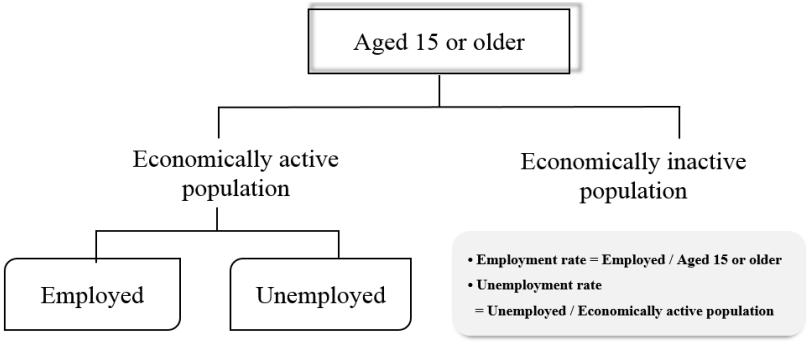
로 나타났다. 이는 NSI가 통계의 속보성과 확장성 등 앞서 설명한 경기판단지표로써 필요한 특성을 고려하였을 때 가계의 경기심리지수로 유용할 수 있음을 보여준다.

<Figure 7~8>은 고용지표인 고용률 및 실업률과 경기심리지수 간의 추이를 비교한 것이다. 직관적으로는 고용률과 심리지수가 동일한 방향으로, 실업률과는 반대 방향으로 움직일 것이라 예상할 수 있다. 소득이 소비로 이어지기까지 일정 시차는 발생할 수 있으나 일반적으로 취업자가 늘수록 경기를 낙관적으로 전망하는 반면 기업의 신규 채용이 감소하거나 해고가 증가할 경우 경기가 좋지 않다고 판단할 것이기 때문이다. 그러나 고용률과 가계의 경기심리지수 간의 추이를 그린 <Figure 7>를 보면 두 지수 모두 고용률과 뚜렷한 양(+)의 상관관계가 보이지 않음을 알 수 있다.



〈Figure 7〉 Employment rate: NSI vs. CSI of trends

왜 이러한 관계를 보이는 것일까? 이는 고용지표의 정의와 관련되어 있다(<Figure 8> 참조). 실업자의 정의에 따르면 15세 이상 인구 중 취업상태에 있지는 않으나 구직활동을 하고 있느냐에 따라 실업자, 즉 경제활동인구로 분류되거나 비경제활동인구로 분류된다(각주 9 참고). 가령 15세 이상 인구와 취업자 수에는 변화가 없다고 가정하고, 고용활동조사에 응답할 당시 취업은 하지 않았지만 구직의사가 있다고 답할 경우 이 사람은 경제활동인구로 분류되어 실업률에는 반영되지만 고용률에는 포함되지 않는다. 반면 동일한 상황에서 구직의사가 없다고 할 경우 실업자가 아니기 때문에 비경제활동인구로 분류되고, 여전히 15세 이상 인구에는 해당하기 때문에 실업률에 반영되지 않고, 고용률에만 영향을 미친다. 이는 고용률과 실업률 간의 상관관계에서도 확인할 수 있는데, 두 지표 간에 강한 음(-)의 상관관계가 아닌 -0.3의 관계를 보인다. 즉 실업자 혹은 실업률이 증가한다고 반드시 고용률이 하락하지 않는다는 것이다. 사람들의 구직의사 결정에 영향을 미치는 가장 큰 요인 중 하나가 경기상황에 대한 판단일 것이므로 경기심리지수는 고용지표에 대해 선행성을 가질 가능성이 높다. 통상 일정기간의 경기상황을 보면서 사람들

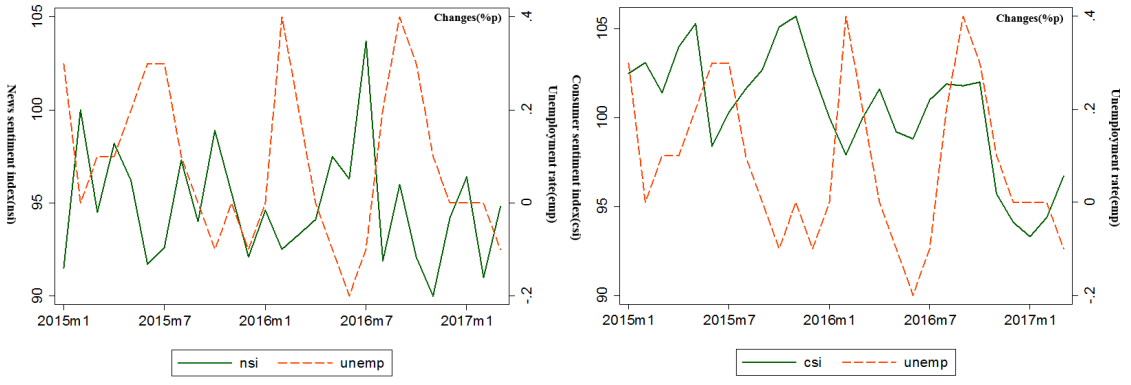


〈Figure 8〉 Composition of Labor indicators

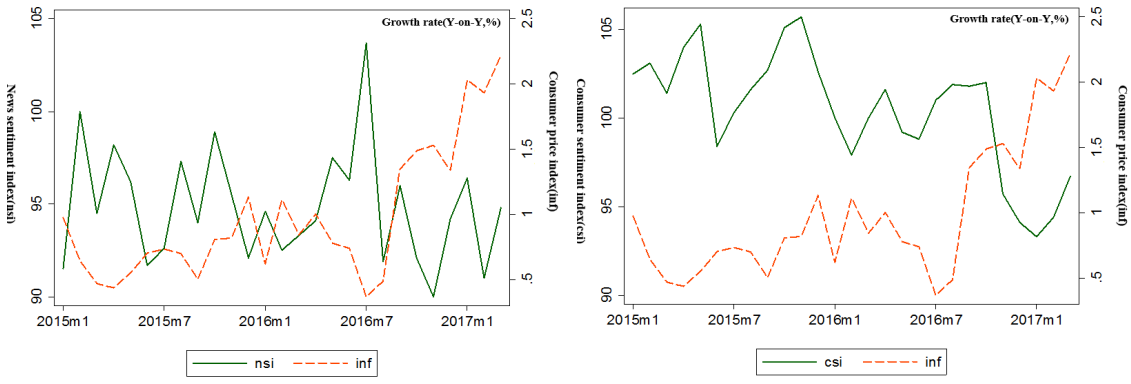
이 구직활동 여부를 결정하기 때문에 경기심리지수가 고용률과 실업률에 반영되기는 시차가 있을 것이기 때문이다. 따라서 동일 시점 간 비교가 중심인 추이분석에서는 이러한 관계를 확인하기 어려우므로 교차상관분석에서 살펴보고자 한다.

한편 실업률과 두 경기심리지수 모두 서로 반대방향으로 움직이는데, 이는 직관적으로도 타당해 보인다. 실업률이 높아질수록 경기 상황을 비관적으로 볼 것이기 때문이다. 또한 실업률의 변동폭이 커서 상대적으로 안정적인 움직임을 보인 CSI에 비해 NSI와 고용지표 간에 보다 강한 음(-)의 상관관계를 보였다. CSI는 0.1, NSI는 -0.4로 나타나 고용동향 파악에 있어 NSI가 보다 나은 정보를 제공하고 있음을 시사한다.

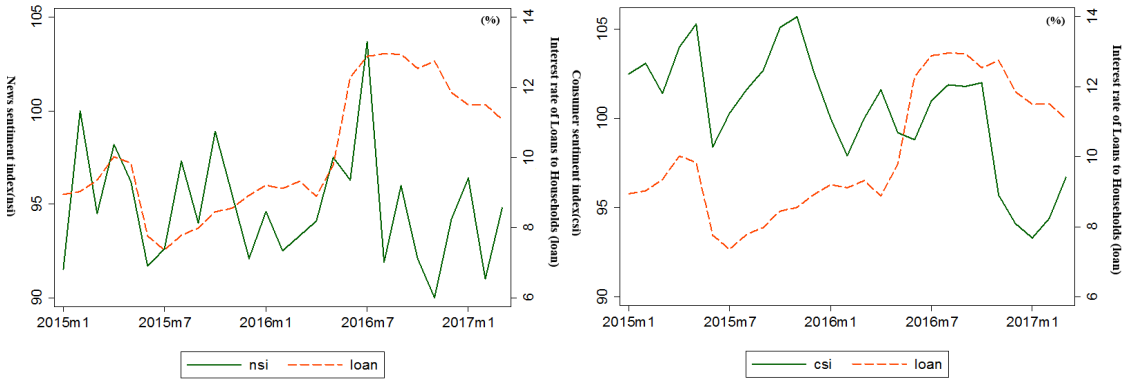
마지막으로 가격지표인 소비자물가상승률과 금리 간의 관계를 분석하였다. 단기적으로 물가상승은 가계의 지출부담을 높이기 때문에 경기인식에 부정적인 영향을 미칠 것으로 예측할 수 있다. 고용지표와는 반대로 소비자물가상승률의 변동 폭이 작기 때문에 실업률과 반대로 NSI에 비해 CSI와의 관련성이 높아 보인다(<Figure 10>).



〈Figure 9〉 Unemployment rate: NSI vs. CSI of trends



〈Figure 10〉 Inflation(Consumer price): NSI vs. CSI of trends



〈Figure 11〉 Interest rate of Loans to Households: NSI vs. CSI of trends

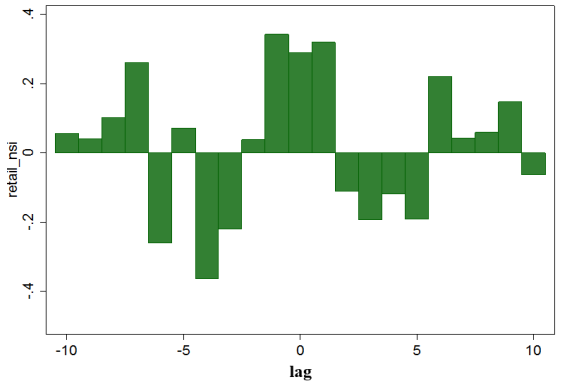
소비자물가상승률과 동일하게 이자율 상승 역시 가계의 이자지출을 늘리므로 가계의 생활형편을 악화시킬 것이기 때문에 두 지표 간에 상반된 관계를 예상할 수 있다. 그러나 기대와 달리 NSI와 가계의 대출금리 간에는 뚜렷한 관계가 보이지 않았다(<Figure 11>). 이는 앞서 소비나 고용활동에 관한 의사를 결정하는 주체가 가계라고 한다면 이와 달리 가계의 대출금리를 결정하는 주체가 가계가 아닌 금융당국이기 때문에 두 변수 간의 관계가 약하게 나타난 것으로 해석된다. 이에 대해서는 교차상관분석에서 자세히 살펴본다.

4.2 교차상관 분석

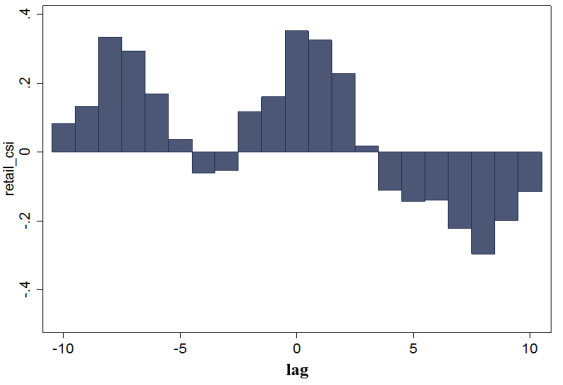
경기심리지수를 통한 경제동향 및 전망을 위해서 다른 거시경제지표들 간의 관계가 중요하다는 것은 앞서도 강조한 바 있다. 경기동행지수는 현재의 경기흐름을 파악하는데 유용하나, 경기의 선행지표가 정책 판단에 있어 활용도가 더 높다. 통계 수집 및 작성에 시간이 소요되어 실제 경제상황과 경기동행지수가 공표되는 시점 사이에는 시차가 발생하기 때문이다. 적시성 있는 정보 제공에 있어 경기동행지수가 가지는 한계이다. 본 연구에서는 교차상관계수의 최대값을 기준으로 지표들과 경기심리지수 간 시차 구조를 확인하였다.

교차상관 분석을 수행한 결과, NSI는 민간소비의 선행지표로 많이 사용되는 CSI와 비교해 소매판매업지수에 대해 양(+)의 방향으로 1개월 선행하면서, 상관관계 또한 더 높은 것으로 나타났다(<Figure 12~13> 참고).

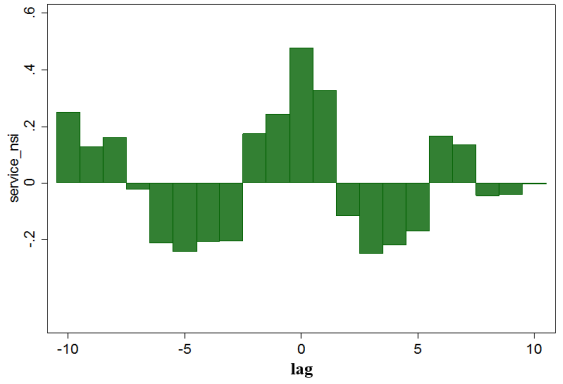
다음으로 서비스업생산지수를 살펴보았다. 서비스는 외식이나 여행과 같이 그 지출의 성격 상 경기에 즉각적으로 반응할 가능성이 높기 때문에 소매판매업지수에 비해 선행성은 약해졌지만 동일시점에 NSI는 높은 상관관계를 보여 서비스에 대한 동행지수로써 유용함이 확인되었다. 반면 CSI는 1개월 후행시점에 높은 상관계수가 나타나 경기판단지수로써 NSI와 비교해 적시성이 떨어지는 것으로 나타났다.



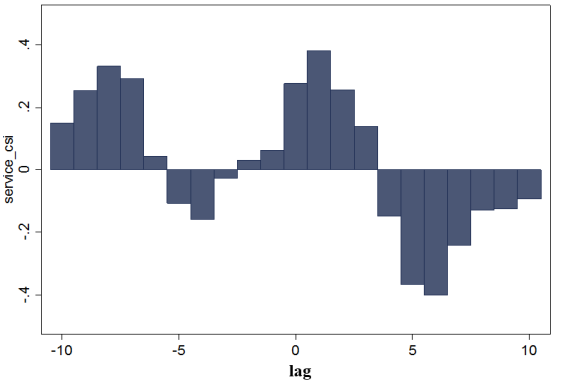
〈Figure 12〉 NSI and Retail sales index



〈Figure 13〉 CSI and Retail sales index



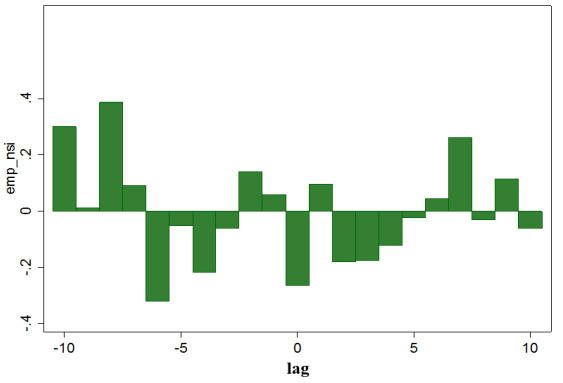
〈Figure 14〉 NSI and Service production index



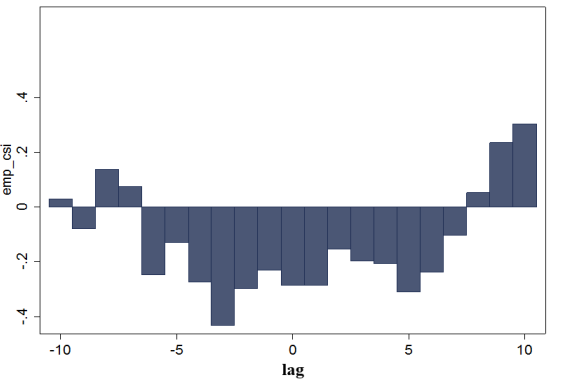
〈Figure 15〉 CSI and Service production index

<Figure 16~17>은 고용률에 대한 교차상관분석 결과이다. NSI는 동일시점 기준으로 8개월 전에 양(+)의 방향으로 상관관계가 높다가 2개월 후에 음(-)의 값으로 전환되었다. 이는 가계가 경기상황에 대한 인식에 따라 경제활동인구에서 비경제활동인구로 2개월 정도의 조정기간을 거치는 것으로 해석할 수 있다. 반면 CSI는 거의 전기간에 걸쳐 음(-)의 상관관계만 보여 가계의 경기판단에 따른 구직의사 결정에 관한 정보가 심리지수에 반영되지 않는 것으로 나타났다.

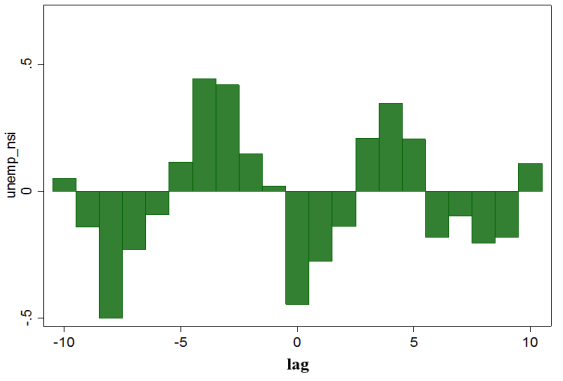
다음은 실업률과 경기심리지수 간 분석이다. 앞서 추이분석 결과를 설명하면서 경기심리지수와 실업률 간에 음(-)의 상관관계를 예상하였다. <Figure 18>를 보면, NSI는 실업률과 8개월 이전에 가장 높은 음(-)의 상관관계를 보이다 4개월 이전부터는 양(+)의 방향으로 전환되었다. 이후 동일시점에 다시 음(-)의 값을 보인다. 이는 고용률과 유사하게 실업으로 인한 경기판단이 8개월 전에 이뤄지며, 실업 상태가 가장 악화된 시점에 비관적으로 경기를 인식하는 것으로 해석할 수



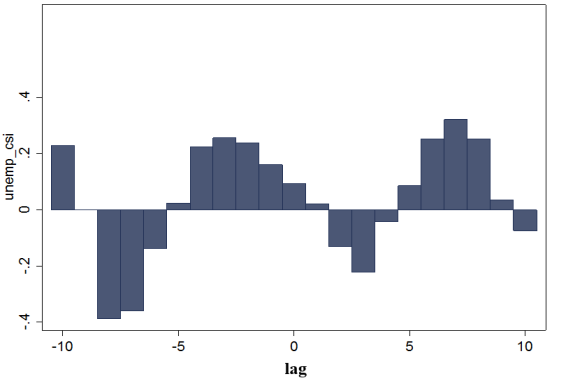
〈Figure 16〉 NSI and Enemployment rate



〈Figure 17〉 CSI and Enemployment rate



〈Figure 18〉 NSI and Unemployment rate

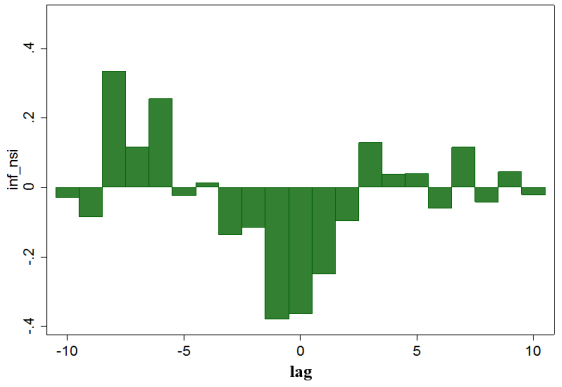


〈Figure 19〉 CSI and Unemployment rate

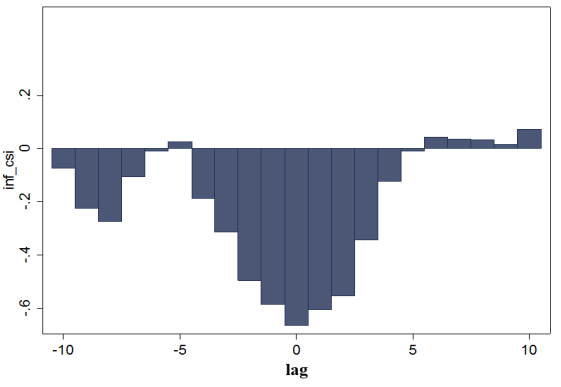
있다. 한편 CSI는 NSI와 유사한 시차구조의 형태를 보이지만 상관관계는 낮아졌다. 고용률과 실업률의 결과를 종합적으로 검토해보면, NSI가 고용동향 및 전망을 판단함에 있어 CSI와 비교해 더 많은 정보를 제공하고 있음을 알 수 있다.

마지막으로 가격지표에 대한 교차상관분석 결과이다. 먼저 소비자물가상승률을 보면, NSI는 1~2개월 음(-)의 방향으로 선행하는 것으로 나타났지만 CSI에 비해 그 정도는 낮았다(<Figure 20~21>). 이는 앞서 언급하였듯이 변수들의 변동폭과 관련이 있다. 소비자물가상승률은 상대적으로 안정적인 움직임을 보이므로 NSI에 비해 CSI와의 관계가 더욱 뚜렷하게 나타났다.

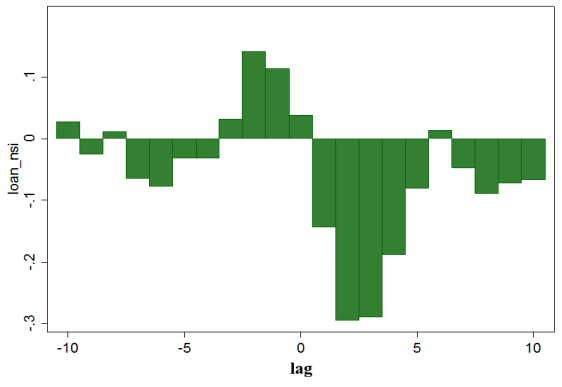
가계의 대출금리에 대해서는 NSI보다 CSI의 시차구조 및 방향성이 보다 합리적인 것으로 나타났다. 경기심리지수는 앞서의 실물지표나 고용지표와 달리 가계의 대출금리에 대해서 후행변수일 가능성이 높다. 금융당국은 경기상황을 주시하며 시장이 지나치게 과열하거나 침체될 것으로 전망되면 금리를 변동시켜 경기를 조절



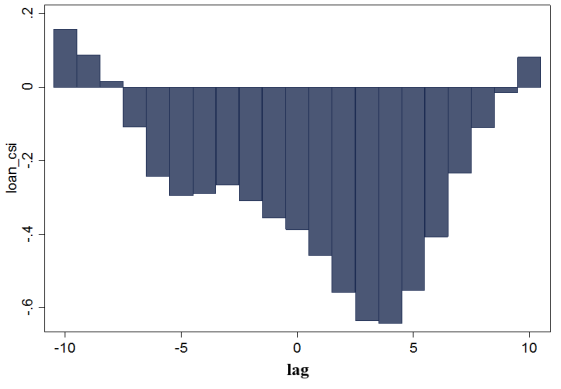
〈Figure 20〉 NSI and Inflation



〈Figure 21〉 CSI and Inflation



〈Figure 22〉 NSI and Interest rate



〈Figure 23〉 CSI and Interest rate

하기 때문이다. 통상 금융당국이 가계보다 경제 상황을 판단할 수 있는 정보를 많이 가지고 있어 가계의 경기판단이 심리지수에 반영되기 전에 금리가 움직일 것이다.

한편, 대출 금리 인상은 가계에는 이자에 대한 지출을 높이기 때문에 경기심리지수와 음(-)의 상관관계를 예상할 수 있다. 이는 NSI와 CSI 모두 동일하게 확인할 수 있었는데, CSI가 3~4개월 후행하는 시점에 -0.6으로, NSI는 -0.3으로 나타났다.

4.3 예측력 분석

본 연구에서 생성한 소비자 경기심리지수의 예측력을 확인하기 위해 단순회귀분석을 수행하였다. 이를 위해 CSI를 설명변수로 설정한 추정식을 기본모형(Base model)으로 설정하였다. 기본모형과 비교하여 본 연구에서 생성한 소비자의 경기심리지수에 대한 예측력을 확인하기 위해 NSI가 설명변수인 추정식을 수정모형(Revised model)으로 하였다.

추정식에서 y는 각각 소매판매업지수 및 서비스업생산지수의 전년동월대비 증가율, 고용률과 실업률의 전년동월대비 증감, 소비자물가의 전년동월 대비 증가율, 가계의 대출금리이다. β는 추정모형에서 예측지표의 계수이며 ε는 오차항이다. γ는 AR과 MA항에 대한 계수에 해당하며, 결과의 해석에서 중요하지 않아 본문의 표에서는 생략하였다. 모형의 시차는 Akaike information criterion (AIC)를 통해 ARMA(1,1)로 설정하였다. 모형의 적합성은 주요 예측 설명변수가 유의하고, R^2가 클수록, F-통계량이 유의할수록 높다고 해석하였다. 선행지표의 예측력은 RMSE(Root Mean Square Error) 및 MAE(Mean Absolute Error)가 낮을수록 향상되었다고 보았다.

Base model: $y_t = \alpha + \beta CSI_t + \gamma_1 y_{t-1} + \gamma_2 \varepsilon_{t-1} + \varepsilon_t$

Revised model: $y_t = \alpha + \beta CSI_t + \gamma_1 y_{t-1} + \gamma_2 \epsilon_{t-1} + \varepsilon_t$

y: 소매판매업지수(retail), 서비스업생산지수(service), 고용률(emp), 실업률(unemp), 소비자물가지수(inf), 가계의 대출금리(loan)
CSI: 소비자심리지수
NSI: 소비자의 경기심리지수
ε : 오차항

먼저 소매판매업지수에 대한 추정결과를 <Table 4>에 정리하였다. 표에서 보듯 NSI와 CSI의 계수는 모두 유의하였지만 F-통계량은 그렇지 않았다. 반면 예측력 면에서는 기본모형이 수정모형에 비해 더 나은 성과를 보였다. 반면 서비스업생산지수에 대해서는 NSI는 1% 수준에서, F-통계량은 5% 수준에서 유의한 것으로 나타나 CSI에 비해 모형의 적합도가 개선되었다. 예측력 또한 RMSE와 MAE가 더 작은 값을 보여 향상되었음이 확인되었다. 다음으로 회귀계수의 부호를 살펴보면, 소매판매업과 서비스업 모두 양(+)의 값으로 나타났다. 이는 가계가 경기를 낙관적으로 인식할수록 소비가 증가함을 의미한다. 앞서 추이 및 교차상관 분석의 해석과도 일치하는 결과이다.

<Table 5>은 노동지표인 고용률과 실업률에 대한 추정결과를 정리한 것이다. 민간소비와 유사하게 기본모형과 비교해 NSI를 예측변수로 사용한 수정모형이 모형의 적합도와 예측력이 개선되어 뉴스기사에서 추출한 정보로 생성된 경기심리지수가 고용시장의 동향 및 전망에 유용한 정보를 제공하는 것으로 확인되었다. 이 역시 교차상관분석과 일관된 결과이다.

〈Table 4〉 Estimation of Real economic indicators

	Retail sales index		Service production index	
	Base model	Revised model	Base model	Revised model
	Coef.(P>|z|)	Coef.(P>|z|)	Coef.(P>|z|)	Coef.(P>|z|)
NSI	- -	0.280*(0.082)	- -	**0.151*****(0.005)**
CSI	0.257*(0.072)	- -	0.058(0.291)	- -
F-stat.(Prob.)	1.866(0.164)	1.353(0.282)	3.316**(0.038)	3.817**(0.023)
R^2	**0.196**	0.150	0.302	**0.332**
RMSE	**2.023**	2.080	0.669	**0.654**
MAE	**1.587**	1.650	0.531	**0.477**

Note: In the equation, AR term, MA term and constant term are included, but those estimation results are omitted in the table. ***, **, and * denote the significance level at 1%, 5% and 10 % respectively.

〈Table 5〉 Estimation of Labor indicators

	Employment rate		Unemployment rate	
	Base model	Revised model	Base model	Revised model
	Coef.(P>|z|)	Coef.(P>|z|)	Coef.(P>|z|)	Coef.(P>|z|)
NSI	- -	**-0.020***(0.009)**	- -	**-0.007**(0.044)**
CSI	-0.015(0.180)	- -	0.009(0.389)	- -
F-stat.(Prob.)	0.794(0.510)	**2.998**(0.051)**	8.200***(0.001)	10.691***(0.000)
R^2	0.094	**0.281**	0.517	**0.582**
RMSE	0.166	**0.148**	0.114	**0.106**
MAE	0.124	**0.120**	0.094	**0.087**

Note: In the equation, AR term, MA term and constant term are included, but those estimation results are omitted in the table. ***, **, and * denote the significance level at 1%, 5% and 10 % respectively.

〈Table 6〉 Estimation of Price indicators

	Consumer price index		Interest rate of Loan to Households	
	Base model	Revised model	Base model	Revised model
	Coef.(P>|z|)	Coef.(P>|z|)	Coef.(P>|z|)	Coef.(P>|z|)
NSI	- -	**-0.029*(0.069)**	- -	0.008(0.405)
CSI	-0.007(0.772)	- -	0.027(0.517)	- -
F-stat.(Prob.)	14.1***(0.000)	21.7***(0.000)	60.8***(0.000)	62.8***(0.000)
R^2	0.648	**0.739**	0.892	**0.895**
RMSE	0.294	**0.252**	0.592	**0.591**
MAE	0.227	**0.204**	0.475	**0.465**

Note: In the equation, AR term, MA term and constant term are included, but those estimation results are omitted in the table. ***, **, and * denote the significance level at 1%, 5% and 10 % respectively.

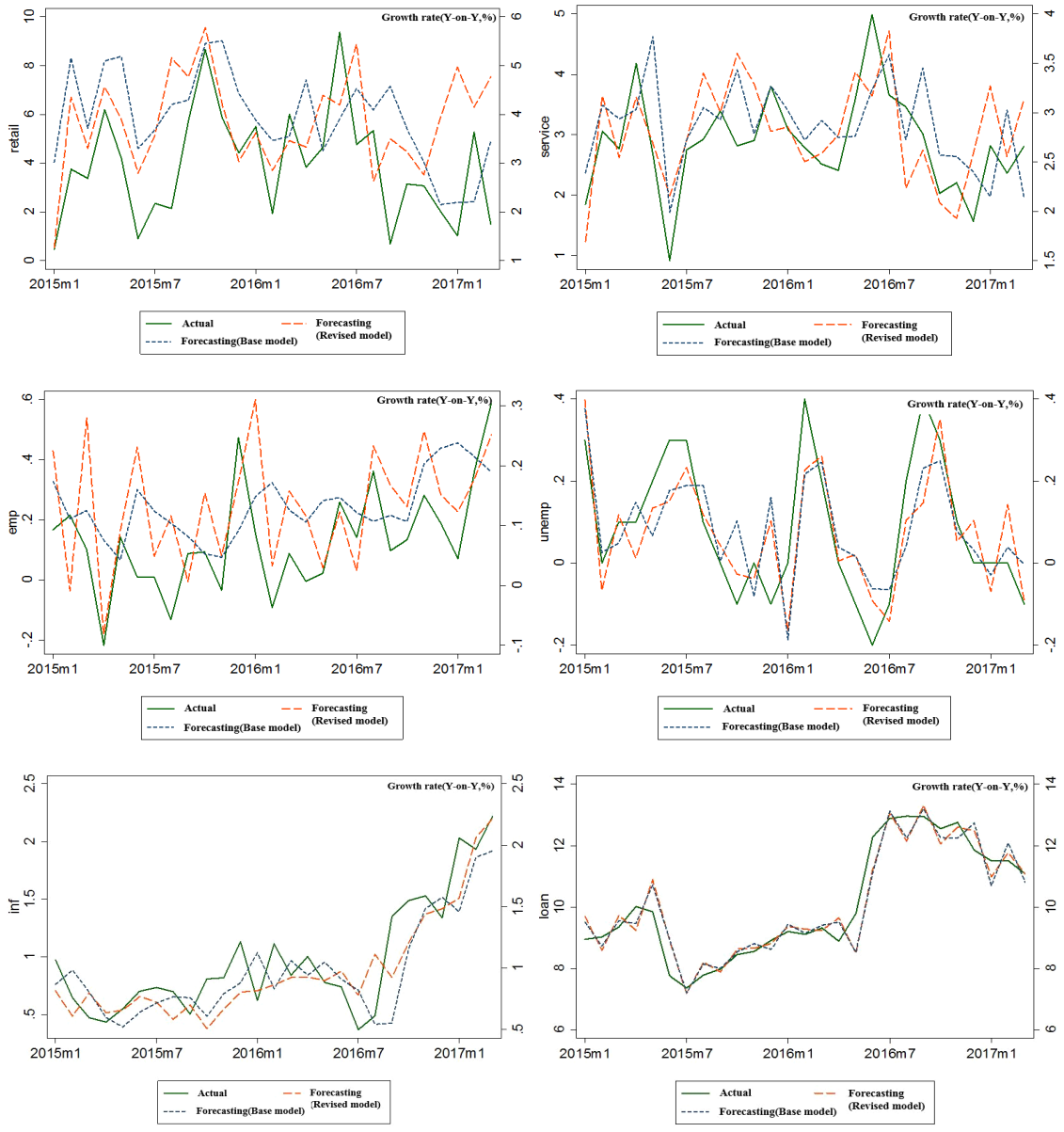
마지막으로 <Table 6>는 가격지표의 추정결과이다. 소비자물가상승률에 대해서 CSI의 계수는 유의하지 않았으나 수정모형인 NSI의 계수는 10% 수준에서 유의한 것으로 나타났으며, 두 모형 모두 F-통계량은 1% 수준에서 유의하였다. 이는 모형의 적합도 면에서는 NSI가 개선되었음을 보여준다. 예측력 역시 수정모형의 성과가 더 좋았다.

반면 가계의 대출금리에 대해서는 두 심리지수 모두 회귀계수가 유의하지 않았는데, 이는 앞서 설명하였듯이 가계의 대출금리가 경기심리지수에 대해 선행 정보를 가지기 때문이다[11]. 한편 추정모형에 포함된 모든 변수의 유의성을 고려한 F-통계량 결과 값은 기본모형과 수정모형 모

11) 본 논문의 연구결과에서는 보고하지 않았지만 그랜저 인과검증을 수행한 결과, 가계의 대출금리는 NSI에 대해서는 1% 수준에서, CSI에 대해 5% 수준에서 유의한 것으로 나타났다. 이는 대출금리의 과거 값이 경기심리지수에 대해 예측력을 가짐을 의미한다.

두 유의하였다. 예측력은 NSI를 예측변수로 사용하였을 때 RMSE와 MAE가 더 작은 값을 가져 분석결과가 향상된 것으로 나타났다.

<Figure 24>는 기본모형과 수정모형에서 추정된 모든 지표의 1 step ahead of out of sample 예측결과를 정리한 것이다. 앞서 단순회귀분석 추



〈Figure 24〉 Out of sample prediction: Actual and Forecasting (Base vs. Revised model)

정결과를 정리한 표에서 보듯이 소매판매업지수는 CSI가, 서비스업생산지수는 NSI를 사용했을 때 실측 값과 더 유사한 추이를 보였다. 그러나 소매판매업지수에 대해서도 실측 값과 비교해 추세나 증가율 수준에 큰 차이가 없어 NSI 또한 소비재의 예측지표로써 유용하게 사용될 수 있을 것으로 보인다.

한편 고용률은 NSI를 사용한 수정모형의 결과가 기본모형의 결과보다 예측력이 확연히 개선된 것으로 나타났다. 실업률과 가격지표에 대한 예측치는 기본모형과 수정모형 간의 큰 차이를 보이지 않았다. 이러한 방식의 예측력 점검은 보다 데이터를 축적하여 분석기간을 늘리고, 분기별 지표 등에도 적용한다면 분석 결과의 견고성(Robustness)을 높일 수 있을 것이다.

5. 결론 및 시사점

경제주체들의 경기상황에 대한 판단 및 전망은 단기적으로 경기변동에 영향을 미치므로 경기심리지수와 거시경제지표들 간에는 밀접한 관련성을 가지는 것으로 알려져 있다. 가계의 경기심리를 나타내는 지표로 국내에서 많이 사용되는 지수에는 설문조사를 통해 작성되는 소비자동향조사가 있다. 그러나 기초자료의 성격 상 경제지표로써 적시성 및 속보성이 떨어지는 문제가 있다. 본 연구에서는 이러한 정형데이터의 한계를 보완할 수 있도록 비정형데이터로 소비자의 경기심리지수를 생성하여 경제분석에서의 활용 가능성을 검토하였다. 민간소비와 관련된 실물지표에는 소매판매업지수와 서비스업생산지수를 사용하였고, 고용지표에는 고용률과 실업률을, 가격지표에는 소비자물가상승률과 가계의 대출금리를 사용하여 가계의 경기심리지수와 지표들 간 추이 및 교차상관분석을 수행하였다. 마지막으로 이들 지표들에 대한 예측 가능성을 분석하였다. 분석결과, 선행지표로 많이 사용되는 소비자심리지수에 비해 높은 상관관계를 보이며, 1 ~2개월 선행한 것으로 나타났다. 예측력 또한 개선된 것으로 나타났다.

경제분석에서 비정형데이터를 활용한 국내연구는 초기단계지만 데이터의 유용성이 확인되면 그 활용도는 크게 높아질 것으로 기대된다. 특히 온라인에서 생성되는 뉴스기사나 소셜 SNS 등의 텍스트 데이터는 속보성이 높고, 커버리지가 넓어 특정 경제적 이슈가 발생할 경우 이것이 경제에 미치는 영향을 빠르게 파악할 수 있다는 점에서 경기판단지표로써의 잠재적 가능성이 클 것으로 보인다.

본 연구에서는 앞서 언급한 몇 가지 문제점 때문에 검색 키워드를 기사의 본문이 아닌 제목에 등장한 경우로 한정하여 분석대상의 범위를 좁혔다. 그러나 보다 정교한 문서 추출법 등을 고려하여 더 많은 기사를 분석에 포함시키는 것이 보다 실제의 경기상황을 반영하는 방법이 될 것이다. 이는 본 연구의 한계이다. 한편 감성분석에서 감성사전의 역할이 매우 중요하기 때문에 영어의 경우 SentiWordNet과 같은 감성사전이 있으며, 그것의 타당성 및 개선 방안에 관한 상당한 선행연구가 축적되어 있다. 반면 한국어는 공개적으로 많이 활용되고 있는 감성사전이 없어 본 연구에서는 별도로 생성했으며, 구축된 감성사전의 객관성 및 타당성에 대한 검증이 필요할 것이다. 그러나 감성사전 구축 방법 자체가 별개 논문으로 다뤄지는 주제이며, 본 논문의 범위를 넘어서므로 추후 연구과제로 남긴다.

참고문헌(References)

A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Goot, M. Halkia, B. Pouliquen, and J. Belyaeva., "Sentiment analysis in the News," *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta, 19-21, May (2010), 2216~2220.

C. C. Aggarwal, and C. Zhai, "Mining Text Data," *Springer Science & Business Media* (2012)

A. Kennedy and D. Inkpen., "Sentiment classification of movie and product reviews using contextual valence shifters," *Computational Intelligence*, Vol.22, No.2 (2006), 110~125.

A. Neviarouskaya, H. Prendinger, and M. Ishizuka., "Sentiful: Generating a reliable lexicon for sentiment analysis," *International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII* (2009), 1~6.

B. Liu., "Sentiment Analysis: Mining Opinions, Sentiments and Emotions," *Cambridge University Press* (2015)

B. Pang, L. Lee and S. Vaithyanathan., "Thumbs up?: sentiment classification using machine learning techniques," in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, *Association for Computational Linguistics*, Vol.10 (2002), 79~86.

B. Pang and L. Lee., "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieva*l Vo1.2. No.1-2 (2008), 1~135.

H. Kanayama and T. Nasukawa., "Fully automatic lexicon expansion for domain oriented sentiment analysis," in: Proceedings of the 2006 Conference on Empirical Methods in *Natural Language Processing* (2006), 355~363.

Hwang, Y.J, "Analysis of consumer propensity using SNS data," *Statistical Research Institute* (2015), 117~155.

J. Bollen, H. Mao and X. Zeng., "Twitter mood predicts the stock market," *Journal of Computational Science*, Vol.2, Issue.1 (2011), 1~8.

J. Bram and S. Ludvigson., "Does Consumer Confidence Forecast Household Expenditure? A Sentiment Index Horse Race," *FRBNY ECONOMIC POLICY REVIEW* (1997)

J. Kamps, M. Marx, R.J. Mokken and M. de Rijke., "Using WordNet to measure semantic orientations of adjectives," *In Proceedings of the International Conference on Language Resources and Evaluation*, Vol.4 (2004), 1115~1118.

Lee, G.A. and Hwang, S.P., "Business Cycle Indicator Using Big Data: Compilation of the Naver Search Business Index," *Economic Analysis*, Vol.20, No.4 (2014), 1~37.

L. Hoang, J.-T. Lee, Y.-I. Song, and H.-C. Rim., "Combining local and global resources for constructing an error-minimized opinion word dictionary," *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence* (2008), 688~697.

M. Hu and B. Liu., "Mining and summarizing customer reviews," in: Proceedings of the Tenth ACM SIGKDD International Conference on *Knowledge Discovery and*

Data Mining (2004), 168~177.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede., "Lexicon-Based Methods for Sentiment Analysis," *Association for Computational Linguistics*, Vol.37, N.2 (2011), 267~307.

N.O., J and K.-s., Shin. "Bankruptcy Prediction Modeling Using Qualitative Information Based on Big Data Analytics," *Journal of intelligence and information systems*, Vol. 22, No.2 (2016), 33~56.

P. D. Turney., "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of Annual Meeting of the Association for computational Linguistics* (2002), 417~424.

P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *Journal ACM Transactions on Information Systems*, Vol.21, Issue.4 (2003), 315~346.

S. Gelpera, A. Lemmensa, and C. Croux., "Consumer sentiment and consumer spending: decomposing the Granger causal relationship in the time domain," *Applied Economics*, 39 (2007) 1~11.

S. Huang, Z. Niua and C. Shi., "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowledge-Based Systems*, Vol.56 (2014), 191~200.

S. Kim and N. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification," *Journal of intelligence and information systems*, Vol. 20, No. 1, (2014), 133~148.

S. Lee, J. Cui, and J. Kim. "Sentiment analysis on movie review through building modified sentiment dictionary by movie genre," *Journal of intelligence and information systems*, Vol. 22, No, 2 (2016), 97~113.

S. Ludvigson., "Consumer confidence and consumer spending," *The Journal of Economic Perspectives*, Vol.18, No. 2 (2004), 29~50.

S. Mohammad, C. Dunne, and B. Dorr., "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus," Proceedings of the 2009 Conference on Empirical Methods in *Natural Language Processing*, Vol.2 (2009), 599~608

TH. Nguyen, K. Shirai and J. Velcin., "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, Vol.42, Issue.24 (2015), 9603~9611.

T. Wilson, J. Wiebe and P. Hoffmann., "Recognizing contextual polarity in phrase level sentiment analysis," in: Proceedings of the Conference on Human Language Technology and Empirical Methods in *Natural Language Processing* (2005), 347~354.

X. Li, H. Xie, L. Chen, J. Wang and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems*, Vol.69 (2014), 14~23.

Y. Lu, M. Castellanos, U. Dayal and C. Zhai., "Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach," in proceedings of *World Wide Web* (2011), 346~356.

Abstract

Construction of Consumer Confidence index based on Sentiment analysis using News articles

Minchae Song* · Kyung-shik Shin**

It is known that the economic sentiment index and macroeconomic indicators are closely related because economic agent's judgment and forecast of the business conditions affect economic fluctuations. For this reason, consumer sentiment or confidence provides steady fodder for business and is treated as an important piece of economic information. In Korea, private consumption accounts and consumer sentiment index highly relevant for both, which is a very important economic indicator for evaluating and forecasting the domestic economic situation. However, despite offering relevant insights into private consumption and GDP, the traditional approach to measuring the consumer confidence based on the survey has several limits. One possible weakness is that it takes considerable time to research, collect, and aggregate the data. If certain urgent issues arise, timely information will not be announced until the end of each month. In addition, the survey only contains information derived from questionnaire items, which means it can be difficult to catch up to the direct effects of newly arising issues. The survey also faces potential declines in response rates and erroneous responses. Therefore, it is necessary to find a way to complement it.

For this purpose, we construct and assess an index designed to measure consumer economic sentiment index using sentiment analysis. Unlike the survey-based measures, our index relies on textual analysis to extract sentiment from economic and financial news articles. In particular, text data such as news articles and SNS are timely and cover a wide range of issues; because such sources can quickly capture the economic impact of specific economic issues, they have great potential as economic indicators.

There exist two main approaches to the automatic extraction of sentiment from a text, we apply the lexicon-based approach, using sentiment lexicon dictionaries of words annotated with the semantic orientations. In creating the sentiment lexicon dictionaries, we enter the semantic orientation of individual

* Big Data Analytics, Ewha Womans University
** Corresponding Author: Kyung-shik Shin
School of Business, Ewha Womans University
52 Ewhayeodae-gil, Seodaemun-Gu, Seoul, 120-750, Korea
Tel: +82-2-3277-2799, Fax +82-2-3277-2776, E-mail: ksshin@ewha.ac.kr

words manually, though we do not attempt a full linguistic analysis (one that involves analysis of word senses or argument structure); this is the limitation of our research and further work in that direction remains possible.

In this study, we generate a time series index of economic sentiment in the news. The construction of the index consists of three broad steps: (1) Collecting a large corpus of economic news articles on the web, (2) Applying lexicon-based methods for sentiment analysis of each article to score the article in terms of sentiment orientation (positive, negative and neutral), and (3) Constructing an economic sentiment index of consumers by aggregating monthly time series for each sentiment word.

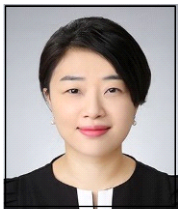
In line with existing scholarly assessments of the relationship between the consumer confidence index and macroeconomic indicators, any new index should be assessed for its usefulness. We examine the new index's usefulness by comparing other economic indicators to the CSI. To check the usefulness of the newly index based on sentiment analysis, trend and cross - correlation analysis are carried out to analyze the relations and lagged structure. Finally, we analyze the forecasting power using the one step ahead of out of sample prediction. As a result, the news sentiment index correlates strongly with related contemporaneous key indicators in almost all experiments. We also find that news sentiment shocks predict future economic activity in most cases. In almost all experiments, the news sentiment index strongly correlates with related contemporaneous key indicators. Furthermore, in most cases, news sentiment shocks predict future economic activity; in head-to-head comparisons, the news sentiment measures outperform survey-based sentiment index as CSI.

Policy makers want to understand consumer or public opinions about existing or proposed policies. Such opinions enable relevant government decision-makers to respond quickly to monitor various web media, SNS, or news articles. Textual data, such as news articles and social networks (Twitter, Facebook and blogs) are generated at high-speeds and cover a wide range of issues; because such sources can quickly capture the economic impact of specific economic issues, they have great potential as economic indicators. Although research using unstructured data in economic analysis is in its early stages, but the utilization of data is expected to greatly increase once its usefulness is confirmed.

Key Words : Sentiment analysis, Unstructured data, Economic sentiment index, Consumer sentiment index, Economic trends and outlook

Received : May 29, 2017 Revised : July 31, 2017 Accepted : September 13, 2017
Publication Type : Regular Paper Corresponding Author : Kyung-shik Shin

저 자 소 개



송 민 채
현재 이화여자대학교 빅데이터 분석학 박사과정에 있다. 이화여자대학교에서 경제학으로 석사 학위를 받고, 산업연구원, 한국은행, 한국개발연구원의 연구원으로 재직했다. 주요 연구분야는 데이터 마이닝과 텍스트 마이닝, 딥러닝, 빅데이터의 경제분석 활용 등이다.

신 경 식
현재 이화여자대학교 경영대학 교수로 재직 중이다. 연세대학교 경영학과를 졸업하고 미국 George Washington University에서 MBA, 한국과학기술원(KAIST)에서 인공지능, 지식기반 시스템 등 지능형 기법을 경영분야에 적용하는 연구로 경영공학 Ph.D.를 취득하였다. 주요 연구분야는 빅데이터 분석 및 인공지능 응용, 데이터 마이닝과 비즈니스 인텔리전스 등이다.