검색 지수를 이용한 거시경제 전망모형 연구†

박성대[1] · 김현학[2]

요약 경제성장률, 물가상승률 같은 거시경제 변수를 예측하기 위한 모형은 정형 데이터만을 사용하여 왔다. 빅데이터 연구에서는 감성지수와 같은 비정형 데이터들이 도입되기도 하였는데, 본 연구에서는 특정한 거시경제변수와 관련된 검색량이 해당 변수에 대한 변동성을 의미하는 것으로 보고 이를 이용하여 거시 경제 예측 모형을 도입하였다. 한편 검색량 지수의 경우, 수집 주기가 매우 빠르다는 점을 활용하여 주기가 다른 다시 거시경제 변수들의 특성들을 활용하여 베이지안 모형 평균 혼합주기 예측모형을 통해 일반적인 예측과 더불어 현재예측 역시 실시하였다. 본 모형을 통해 우리나라 경제성장률 예측을 산업생산지수, 물가, 통화, 유가, 금리, 노동, 임금, 무역 등을 나타내는 변수들로 실시한 결과 검색량 그 자체는 비록 정확한 경제 정보를 반영하는 것은 아니지만, 기존의 거시경제 데이터가 추적하지 못하는 여론의 흐름 혹은 소비자들의 경제심리 등을 반영할 수 있어 기존 경제 모형에 보조적인 변수로 활용될 수 있어 예측력 향상을 기대할 수 있을 것으로 보인다. 매우 단순한 형태의 거시경제 예측모형에서도 검색 결과 지수는 예측력을 향상시키는 결과를 보여 이를 일반균형 모형 등으로 확장하게 되면 상당한 이점이 있을 것으로 보인다.

핵심주제어 비정형데이터, 현재예측, 검색지수, 전망모형
주제분류 E31, E52, C32

---

* 1) 국민대학교 경제학과 학사(psdae62@gmail.com)
2) 교신 저자, 국민대학교 경제학과 부교수(hyunhak.kim@kookmin.ac.kr)

---

† 이 논문은 2022년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022S1A5A2A01044998). 또한 본 논문은 대외경제정책연구원 연구보고서 23-24 “빅데이터 기반의 국제거시경제 전망모형 개발 연구”중 저자들이 집필한 제 4장 “ 비정형 데이터를 이용한 거시경제 전망모형”을 기반으로 함

Ⅰ. 도입

Choi and Varian (2012)의 선구적인 연구 이후, 인터넷 검색 데이터를 이용한 현재예측(nowcasting)은 시의성 있는 거시경제 지표를 찾는 정책 입안자들 사이에서 점점 인기를 얻고 있다. 그러나 검색 엔진 변수의 사용은 주로 농산물 소비(Jang and Choe, 2016), 채소 가격(Yoo, 2016), 자동차 판매량(Carri`ere-Swallow and Labb´e, 2011)과 같은 특정 거시경제 변수를 예측하는 데 주로 사용되었다. Koop and Onorante(2019)는 "구글 검색이 '집단지성'을 수집할 수 있고, 직접적으로 또는 기대를 통해 결과에 영향을 미침으로써 다양한 시점의 모델에서 어떤 거시변수가 중요한지에 대한 정보를 제공할 수 있다는 가설"을 통해 일반적인 거시경제 변수를 예측할 때 구글 트렌드 데이터가 유용하게 사용될 수 있음을 보였다. 이들은 특정 형식에 맞는 예측모형을 선택할 때 구글 검색 데이터를 기존 거시 변수들과 함께 사용하면 실업률, 기간 스프레드, 인플레이션 등과 같은 거시경제 전체적인 예측이 개선되고, 단순히 구글 검색 지수를 이용한 변수를 포함하면 특정 변수의 예측만 개선된다는 것을 보여주었다.

본 연구에서는 한국과 같은 소규모 개방경제의 경우, Koop and Onorante(2019)의 방법론을 통해 거시경제 총량지수를 현재예측(nowcasting) 할 수 있는 성능을 개선할 수 있는지에 대한 문제를 다루고자 한다. 사실 우리나라의 경우 무역규모가 GDP의 80%를 넘어가는 경제임을 감안하면 국내 검색엔진의 검색 지수를 이용하여 우리나라 거시경제 총량변수를 직접적으로 예측하는 것은 효과가 적을 수도 있다. 즉, 한국의 거시경제 총량지수들은 세계 경제뿐만 아니라 국내 경제의 영향을 받기 때문에 국내 검색엔진인 네이버의 변수로 예측하기에는 부적절할 수 있다는 것이다. 하지만 이와 같은 지수들은 단독 변수가 아니라 보조 지표로서 활용한다면 기존의 예측모형의 성능을 더 향상시킬 수 있을 것으로 예상한다. 예를 들어 Lee와 Hwang(2014)은 네이버 트렌드 데이터에서 추출한 대체 경기 및 소비심리지수를 구축하는 데 성공했는데, 이는 Vosen과 Schmidt(2011)에 이어 설문조사에 기반한 지수를 능가하는 성과를 거두는 등, 검색 엔진의 결과가 전반적인 총량지수들의 변화를 추적하는 데 유용할 것으로 보인다.

본 연구는 Koop and Onorante(2019)의 방법론을 이용하여 검색엔진의 특정 단어에 대한 상대적인 검색량을 일종의 확률로서 전환하여 한국 거시경제 총량지수들의 현재시점 예측 성능을 향상시킬 수 있음을 검증하여 문헌에 기여하고자 한다. 다만 해외 시장 노출도가 높은 한국 경제의 특수성 외에도 검색엔진 변수의 단순 포함이 거시경제 총량 예측에 부적절하게 작용하는 다른 문제들이 있을 수 있다. 먼저 네이버가 국내 검색엔진 시장의 지배적 사업자이기는 하지만, 해외 정보를 수집하고 해외 시장에 참여하는 입장에 있는 사람들은 구글을 포함한 다양한 영어 기반의 정보원에 의존하는 경향이 있다는 점이다. 즉, 한국 경제가 국제적 충격에 노출되어 있는 반면, 현재 네이버 검색어는 국내 관심사에 국한되어 있는 경향이 있다. 반면 구글은 국내 검색엔진 시장에서 차지하는 비중이 미미하기 때문에 구글코리아의 트렌드 데이터 역시 변동성이 너무 커서 정보성이 높다고 보기 어렵다.[1] 이런 점에서 네이버 검색어를 한국 거시경제 총량의 예측

1) 다만 이와 같은 검색엔진의 시장 지배력은 시간에 따라 네이버의 시장점유율은 하락하고, 구글외에도 소셜미디어나 유튜브와 같은 전혀 다른 매체를 통해 검색을 하는 등 검색엔진 시장에 상당한

변수로 간주하기보다는 네이버 검색어가 한국 경제의 국면에 대한 집단적 지식을 제공한다고 가정하는 것이 더 타당해 보인다. 검색어 활용과 관련된 또 다른 문제는 검색어의 우리말 표현인데, 이 문제는 네이버에서 제공하는 '카테고리화' 기능으로 어느 정도 해결할 수 있다. 네이버는 2008년 구글 트렌드에 이어 네이버 트렌드 네이밍을 시작했는데, 구글과 마찬가지로 검색 결과는 검색 시간에 비례하고, 그 결과 수치는 전체 주제에 대한 전체 검색에서 해당 주제가 차지하는 비율에 따라 0~100의 범위에서 스케일링된다. 하지만 한국어는 영어와 달리 명제 단위가 단어의 일부가 되는 경우가 많고, 사용 빈도에서 동의어와 동음이의어의 존재로 인해 검색량에서 트렌드를 추출하는 데 문제가 있다[2]. 네이버는 2016년 네이버 트렌드를 네이버 데이터랩(이하 NDL)으로 명칭을 변경한 후, 검색어를 주제별로 분류해 단어가 아닌 주제별로 트렌드 데이터를 추출할 수 있도록 했다. 즉, 연관성 있는 용어는 포함하고 관련 없는 용어는 걸러냄으로써 트렌드 데이터가 이전보다 실제 주제를 더 잘 반영하고 있다. 로컬 검색 엔진의 데이터를 통해 국내 거시경제 지표의 현재 예측 성능이 개선된다면, 정책 입안자들이 경제 문제를 보다 정확하게 감지하고 정책 수단을 적시에 고려하는 데 도움이 될 수 있다. 많은 거시경제 변수들이 시차를 두고 수집되기 때문에 정책 입안자들은 최근 동향에 대한 정보를 뉴스나 설문조사, 일부 시장 참여자들의 추측에 의존하는 경우가 많다. 예측 오차를 줄이면 부정확한 정보에 대한 의존도를 낮출 수 있고, 보다 자신감 있게 정책을 수행할 수 있다.

Ⅱ. 전망모형

1. 모형의 형태

각 모델은 거시경제 변수 중 하나를 종속변수인 y_t로 사용하고 나머지 거시경제 변수는 잠재적 설명 변수인 X_t로 포함한다. X_t에 해당하는 Google 변수는 Z_t라는 레이블이 붙는다. Google 변수는 매주 집계되는 반면, 거시경제 변수는 매월 집계가 된다. 변수는 월별로 사용할 수 있는데, 연구실험에서는 t월 마지막 주의 Google 데이터를 사용하므로 Z_t는 t월 마지막 주에 해당되는 데이터이다. 물론 원하는 시점에 따라 다른 타이밍 규칙을 사용할 수 있는데, 분기-월 데이터로도 전환할 수 있다.

y_t를 예측하기 위한 표준 한 단계 앞선 회귀 모형은 다음과 같다:

$$y_t = X'_{t-1}\beta + \varepsilon \tag{1}$$

일반적으로 이 모형에는 종속 변수의 시차항과 상수항도 포함하지만, 간결성을 위해 생략하였다. 추후 제시되는 모형에서도 기본적으로 시차항과 상수항 역시 포함되지만 식의 간결성을 위해 역

변화가 진행되고 있다. 본 연구에서는 이와 같은 새로운 트렌드를 반영하기에는 자료가 충분치 않기 때문에 여전히 네이버를 압도적인 시장지배자로서 가정하고 네이버 검색엔진의 결과만을 이용한다.

2) 예를 들어 '물가상승률', '인플레이션', '인플레' 는 모두 영어로는 inflation을 지칭하는 말이며, 일상적으로 세 단어 모두 통용된다. 또한 '보수'는 임금을 뜻하기도 하지만 보수주의나 보수정당을 뜻하기도 하는 등, 영어보다 동의어 및 동음이의어가 훨씬 많다.

시 생략되었음을 미리 알려둔다.

식 (1)에 다음 Google 변수라고 하는 검색엔지의 검색빈도의 지수를 추가한다. 이를 위해서 다음과 같은 시간 규칙을 가정한다. t월 말 또는 $t+1$월 초에는 y_t가 관찰되지 않았으므로 현재예측에 관심이 있다고 가정하자. t월의 마지막 주인 Z_t에 대한 Google 검색 데이터를 사용할 수 있게 된다. 다른 거시 경제 변수는 시차를 두고 발표되므로 X_{t-1}은 사용할 수 있지만 X_t는 사용할 수 없는 것으로 가정한다. 이러한 가정을 바탕으로 다음과 같이 $t+1$월 초에 y_t를 현재 예측(nowcasting)할 수 있다.

$$y_t = X'_{t-1}\beta + Z'_t\gamma + \varepsilon \tag{2}$$

수식 (2)는 월말을 기준으로 작성되었지만, 이를 약간만 변경하면 다른 시기(예를 들어 매달 15일 시점에 예측)를 기준으로 작성할 수도 있다. 여기서 가장 중요한 건 거시 변수들, X_t 의 변수 작성일에 따라 다르다.

한편 식 (2)의 예측 프레임워크는 시간이 지남에 따라 계수가 일정하다고 가정하기 때문에 강건성을 유지하기 쉽지 않고 (Stock and Watson, 1996), (잠재적으로) 많은 설명 변수와 (상대적으로) 짧은 시간 범위로 인해 매개변수가 과도하게 설정되어 있다는 단점이 있다. 이런 단점을 극복하기 위해 시간에 따라 변하는 매개변수를 허용하여 거시경제학에서 널리 사용되는 시변모수(Time Varying Parameter, TVP) 회귀모형을 사용할 수 있다.

Koop and Onorante(2019)가 제시한 TVP 회귀 모형은 다음과 같다.

$$y_t = W'_t\theta_t + \varepsilon_t \tag{3}$$

$$\theta_{t+1} = \theta_t + \eta_t \tag{4}$$

여기서 W_t는 상수항뿐만 아니라 X_{t-1}, Z_t 또는 둘 다를 포함할 수 있다. 즉, W_t는 t 시점에 현재 예측 또는 예측에 사용할 수 있는 모든 정보를 나타낸다. 예측 실험에서 시계열 데이터의 빈도는 월별이므로 W_t는 가능한 모든 월별 정보를 나타낸다. 또한, η_t는 i.i.d 를 따른다. 위의 식 (3)은 일종의 '상태 공간' 모형의 전형적인 형태로 칼만 필터를 사용하여 쉽게 추정할 수 있다. 다만 이와 같은 방법의 한 가지 단점은 과도한 매개변수화이다.

이와 같은 매개변수가 너무 많은 문제를 극복하기 위해, 많은 문헌에서 시간에 따른 모수, 즉 시변 모수를 도입하여 모형 평균화나 모형 선택 방법을 사용합니다. 이와 같은 방법의 목적은 보다 (모수가 적은) 작은 모형들을 여러개 모아서 평균화하거나 이와 같은 모형들 중에 선택함으로써 모형이 보다 간결해지는 것이다. 즉, 이를 '축소(shrink)' 모형이라고도 한다. Koop and Korobilis(2012)는 동적 방식으로 TVP 회귀모형을 도입하였고, 이를 통해 동적 모형 평균화 및 선택(DMA 및 DMS) 방법론을 제시하였다. 이러한 모형을 사용하면 시간에 따라 변화하는 가중치를 사용하여 다른 모델을 평균화하거나 각 관측 시마다 다른 모델을 선택할 수 있다. 식 (3)에서 $j=1,...,J$ 로 인덱싱 하면 다음과 같은 다양한 모델을 식별할 수 있다.

$$y_t = W'^{(j)}_t\theta^{(j)}_t + \varepsilon^{(j)}_t \tag{5}$$

$$\theta^{(j)}_{t+1} = \theta^{(j)}_t + \eta^{(j)}_t \tag{6}$$

여기서 $\varepsilon^{(j)}_t$는 $N(0,\sigma^{2(j)}_t)$, $\eta^{(j)}_t$는 $N(0,Q^{2(j)}_t)$를 따른다. $W^{(j)}_t$는 잠재적 설명 변수의 sub-set으로 잠재적인 설명 변수들의 모두 조합을 가리킨다고 할 수 있다. S를 설명 변수의 수라고 하면

총 조합 가능한 모형의 수는 2^S가 된다. 즉, 설명 변수의 후보가 10개인 경우라면 1024개의 모델을 고려해야 하므로 엄청난 계산 부담이 발생할 수 있다. 이와 같은 계산 부담을 해결하기 위한 방법은 모형의 업데이트 방정식을 도출하여 해결하는데 이는 다음 페이지에 다시 설명하고자 한다.

일단 $\sigma_t^{2(j)}$는 지수 가중 이동 평균(EWMA) 방법을 사용하는데 이는 GARCH의 특별한 경우라고 볼 수 있다.

$$\hat{\sigma}_t = \kappa\hat{\sigma}_{t-1} + (1-\kappa)\hat{\varepsilon}_t\hat{\varepsilon}'_t \qquad (7)$$

여기서 $\hat{\varepsilon}$은 추정된 회귀 오차이며, 감쇠 계수 κ는 Riskmetrics(1996)을 따라 0.96으로 설정했다. 또한 상태공간 모형에서 사용되는 소위 '망각계수 (forgetting factor)' 방법을 사용하여 $Q_t^{2(j)}$를 추정한다. EMWA 방법에 의한 추정에 대한 자세한 내용은 Raftery, Karny and and Ettler(2010) 및 West and Harrison(1997)을 참고하기 바란다. 마지막으로 매개변수 $\lambda \in [0, 1]$의 선택이 $\theta_t^{(j)}$에 가중치를 부여하여 θ자체를 추정하도록 한다. 1에 가까운 λ 값을 선택하는 것이 일반적으로, Raftery, Karny and Ettler(2010)은 λ = 0.99로 설정했고, Koop and Korobilis(2012)는 실증 분석에서 모수값을 변화시키면서 민감도를 테스트했다. λ = 0.99라는 것은 분기별 거시경제 데이터의 경우 5년 전 관측치가 직전 기간 관측치보다 약 80%의 가중치를 받는다는 것을 의미한다. 우리나라의 경우 다른 선진국의 사례에 비해 변동성이 더 높을 수 있기 때문에 λ = 0.95, 0.97 및 0.99를 시도하여 매개변수의 민감도를 테스트하고자 한다. 일단 $\sigma_t^{2(j)}$, $Q_t^{2(j)}$를 EWMA와 망각계수를 사용하여 추정하면 다음과 같은 예측 밀도, $p_j(y_t|W_{1:t}, y_{1:t-1})$를 구할 수 있다. 여기서 $W_{1:t}$는 $(W_1,\ldots,W_t)$를, $y_{1:t-1}$은 $(y_1,\ldots,y_{t-1})$를 의미하고 상태공간모형에서 칼만 필터링을 이용해 예측밀도를 추정한다.

DMA 및 DMS에는 $q_{t|t-1,j}$를 이용하여 재귀적 업데이트 방식이 사용된다. $q_{t|t-1,j}$는 모형 j가 시점 t에 $t-1$시점까지의 자료만을 가지고 y_t를 예측 (혹은 현재 예측)할 때 사용될 확률을 의미한다. 그런 다음 $q_{t|t,j}$는 t 시점의 정보를 사용하여 $q_{t|t-1,j}$를 업데이트한다. DMS는 $q_{t|t-1,j}$의 값이 가장 높은 단일 모형을 선택하는 것으로 선택된 모형이 y_t를 예측 (혹은 현재 예측)하는데 사용된다. 따라서 DMS에는 특정한 모형은 없게 되고, 모형이 결국 시간에 따라 계속 바뀌게 된다. DMA는 모든 J 개의 모형에서 시행한 예측을 가중 평균화하는 것으로 그 가중치는 $q_{t|t-1,j}$가 된다. 가중치 역시 시변하므로 이름에서 알 수 있듯이 DMA 역시 동적인 방법론이다.

Raftery, Karny and Ettler(2010)는 다음과 같이 모형의 업데이트 방정식을 도출하였다.

$$q_{t|t,j} = \frac{q_{t|t-1,j}p_j(y_t|W_{1:t}, y_{1:t-1})}{\Sigma_{l=1}^{J} q_{t|t-1,l}p_l(y_t|W_{1:t}, y_{1:t-1})} \qquad (8)$$

여기서 $p_j(y_t|W_{1:t}, y_{1:t-1})$는 모델 j에 대한 칼만 필터에 의한 y_t에 대한 예측 밀도이다. 그런 다음 Raftery, Karny and Ettler(2010)와 같이 망각계수 α를 사용하여 식 (8)을 추정한다. 이는 다음과 같은 모델 예측 방정식을 생성한다.

$$q_{t|t-1,j} = \frac{q_{t-1|t-1,j}^{\alpha}}{\Sigma_{l=1}^{J} q_{t-1|t-1,l}^{\alpha}} \tag{9}$$

여기서 α ∈ (0, 1]이고, 또 다른 망각계수인 α는 1보다 약간 작은 고정 값으로 설정되고, λ와 비슷한 방식으로 해석할 수 있다. 실증 분석에서는 추정의 민감도를 α = 0.95, 0.97 및 0.99로 설정하여 민감도를 체크하고자 한다. 만약 α = 1이면, $q_{t|t-1,j}$는 $t-1$시점까지의 데이터를 사용한 모형에서의 한계 우도값에 비례하게 된다. 이와 같은 방법은 소위 말하는 베이지안 모형 평균법(Bayesian Model Averaging, BMA)이 사용하는 방식이다. 즉, DMA는 BMA의 방법론에 시변계수를 도입하여 동적 요소를 확장한 버전이라고 생각할 수 있다.

이와 같은 방법에서처럼 망각계수 α를 사용하게 되면 상태 공간 안의 모형간의 전이 확률을 유도해야 하는 MCMC 알고리즘에 비해 훨씬 간단한 칼만 필터링을 통해 방정식마다 업데이트 하는 방법을 통해 모형을 추정할 수 있다. 즉, 다음과 같이 모형 업데이트 방정식을 세울 수 있다.

$$q_{t|t,j} = \frac{q_{t|t-1,j} p_j(y_t|y_{1:t-1})}{\Sigma_{l=1}^{J} q_{t|t-1,l} p_l(y_t|y_{1:t-1})} \tag{10}$$

여기서 $p_j(y_t|y_{1:t-1})$는 y_t에서 평가된 모형 j에 대한 예측 밀도이다.

모든 모형에서의 예측 결과를 확률 $q_{t|t-1,j}$를 이용하여 가중 평균하여 재귀적 예측을 실시할 수 있다. 따라서, DMA의 점 예측은 다음과 같의 정의된다.

$$E(y_t|y_{1:t-1}) = \sum_{l=1}^{J} q_{t|t-1,l} z_t^{(l)} \hat{\theta}_{t-1}^{(l)} \tag{11}$$

DMS는 각 시점에서 $q_{t|t-1,j}$의 값이 가장 높은 단일 모형을 선택하여 이 모형의 예측값이 사용된다. 식 (9)의 초기값은 $q_{0|0,j} = \frac{1}{J}$으로 모든 $j = 1,..,J$에 적용할 수 있다. 그리고 나서 두 가지 핵심 요소인 $q_{t|t,j}$와 $q_{t|t-1,j}$를 모든 j에 대해서 업데이트 할 수 있다.

2. 네이버 검색 지수를 이용한 DMA와 DMS

거시경제 총량변수 예측을 위해 네이버 검색 결과의 검색 수치를 직접 설명변수로 고려하는 것은 컴퓨터공학 혹은 데이터 사이언스 문헌에서 여러 번 시도된 적이 있고 실제로도 좋은 예측력을 보인 경우들이 있다. 하지만 보통 그런 경우 과적합의 문제로 경제가 큰 구조 변화를 겪고 나면 해당 모형의 유용성이 떨어지는 경우가 있다. 본 연구에서는 검색엔진의 검색 결과를 기존의 거시경제 변수가 포집할 수 없는 정보를 보충하는 용도로 사용하고자 한다. 검색량이라는 것은 정확하고 부호화된 인과관계가 아니라 특정 변수가 특정 시점의 현재 예측과 관련성이 있다는 것을 보여줄 수 있다는 것이 기본적인 아이디어이다. 따라서 회귀분석에서 설명변수로써 직접적

인 예측력이 거의 없는 검색량이라도 특정 시점의 예측에 가장 기여도가 높은 다른 설명변수를 선정하는 데는 유용할 수 있다. Koop and Onorante(2019)을 따라 이러한 고려 사항들을 바탕으로 앞서 설명한 DMA/DMS를 개선하여 다음과 같은 모형을 제안한다.

먼저 $Z_t = (Z_{1t}, \ldots, Z_{kt})'$를 k개의 검색량 변수로 정의하고, 우리가 고려하고 있는 거시변수들과 매치가 되도록 관련 변수에 관한 검색량을 수집하였다. 각 거시변수와 검색어의 매치업은 부록에 수록하였다. 다시 Z_{it}를 검색 결과를 표준화하여 0과 100사이의 값을 가진 변수라고 가정하자. Koop and Onorante(2019)에서 설명한 것처럼 이 값은 마치 확률과 같이 볼수도 있어 그들은 구글 확률변수로 정의하였다. 우리도 이와 같은 작명을 그대로 따른다면 네이버 확률변수로 정의할 수 있을 것이다.

식 (5)에서 정의한 것과 동일한 정의역에서 $W_t = X_{t-1}$로 정의하자. 각 모형에 대해 시간 t에 대해 우리는 네이버 확률변수라고 부르는 $p_{t,j}$를 다음과 같이 정의하자.

$$p_{t,j} = \prod_{s \in I^j} Z_{st} \prod_{s \in I^{-j}} (1 - Z_{st}) \tag{12}$$

여기서 I^j는 모형 j에 어떤 변수가 있는지를 나타낸다. 예를 들어, 모델 j가 두 번째 및 다섯 번째 설명 변수의 시차항을 포함하는 TVP 회귀 모델이라면, $I^j = 2,5$가 된다. 같은 방법으로 모형 j에서 제외된 설명변수들의 집합은 I^{-j}로 표현한다. $\sum_{j=1}^{J} p_{t,j} = 1$이며, 각 네이버 확률변수는 검색량에

따라 변동하는 것으로 해설할 수 있다. 앞선 예제처럼 $I^j = 2,5$인 상황에서 $p_{t,j}$는 두 변수에 대한 검색량이 많아지면 상승하게 되고, 반대로 해당 변수에 관련한 검색이 줄어들면 수치가 낮아지게 된다. Koop and Onorante(2019)는 Raftery, Karny and Ettler(2010)의 알고리즘을 이용하여 네이버 확률을 반영할 수 있도록 다음과 같이 계량하였다.

$$q_{t|t-1,j} = \omega \frac{q_{t-1|t-1,j}^{\alpha}}{\Sigma_{l=1}^{J} q_{t-1|t-1,l}^{\alpha}} + (1-\omega) p_{t,j} \tag{13}$$

여기서 ω는 0과 1사이의 값으로 연구자가 변경할 수 있는 값이다. 즉, ω가 1이면 이는 Raftery, Karny and Ettler(2010)이 제시한 전통적인 DMA와 DMS와 동일하게 된다. 만약 ω가 0이라면, $q_{t|t-1,j} = p_{t,j}$가 되어 네이버 확률변수가 유일한 설명변수가 되게 된다. 실증분석에서는 다양한 ω값을 시도하여 결과를 살펴보고자 한다.

Ⅲ. 모형 추정 및 전망 예측

1. 데이터

본 연구에서 예측 대상은 우리나라의 GDP만으로 한정하였지만, GDP를 예측하기 위한 거시 총량 자료에는 표 1에서 제시된 바와 같이 여러 분야의 거시경제 변수를 고려하였다. 거시 자료들은 대부분 2000년 이후로는 모두 입수가능하며, GDP와 주기를 맞추기 위해 월간 자료**의** 총량자료는 분기의 합으로 지표는 평균을 기준으로 분기별 자료로 수정되었다.

<표 1> 모형에 사용한 변수 목록

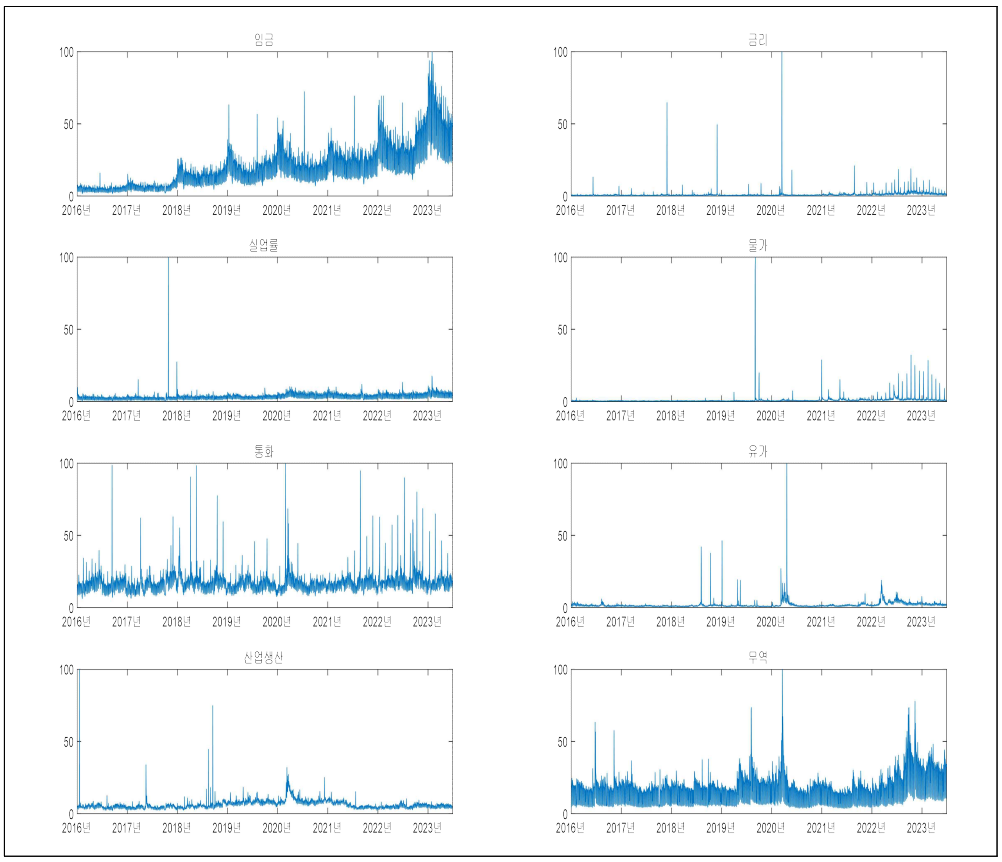
분야	변수	변수변환
종속 변수	GDP	1차 로그 차분
산업	산업생산지수(건설업 제외)	1차 로그 차분
물가	소비자물가지수 (core)	1차 로그 차분
통화	M2	1차 로그 차분
유가	크루드 유가 종합	1차 로그 차분
금리	10년채-콜 금리	변환 없음
노동	실업률	변환 없음
임금	시간당 평균 임금(전체)	1차 로그 차분
무역	순수출(금액기준)	1차 로그 차분

자료: 한국은행(검색일: 2023.9.9.), 통계청(2023.9.9.)

네이버 검색자료는 네이버가 제공하는 네이버 데이터 랩(이하 NDL)에서 직접 수집하였다. NDL에서 주제별 지표를 수집하기 위해서는 검색어가 필요한데, 본 연구에서는 Scott and Varian (2013)을 따라 임의적인 판단(arbitrary judgement)를 가장 줄일 수 있도록 검색어를 설정하였다. 먼저 고려한 거시경제 변수를 기준으로 검색을 실시하고, 네이버가 제시하는 관련 검색어를 추가로 수집하였다. 이와 같은 작업을 반복하면 해당 변수와 관련된 검색어 리스트를 만들 수 있다. 하지만 한국어의 특성상 말을 줄이거나 약자를, 혹은 '인플레'와 같이 영단어를 그대로 사용하는 경우도 있기 때문에 이를 구별하는 과정에서는 임의적인 판단이 개입할 수 밖에 없다. NDL에서는 여런 단어를 동시에 고려하여 검색량을 알려주는 기능이 있어, '물가상승률'과 '인플레이션'을 개별적으로 수집하지 않고 동시에 수집하도록 할 수 있게 한다. 이와 같은 기능을 이용하여 물가와 관련된 모든 단어를 수집하고 이들 단어들의 검색량을 한꺼번에 수집할 수 있도록 하였다.

두 번째로, 400여개에 달하는 검색 단어들을 각 거시 변수에 매치가 될 수 있도록 주성분 분석을 통해 카테고리화하였다. 각 검색어에 따른 검색어의 절대 수는 개인들의 선호나 언어 습관들이 반영되어 있을 수 있기 때문에 이들을 적절히 조합할 수 있는 방법은 없다. 그렇기 때문에 주성분 분석을 통해 변수들의 공통성을 가장 잘 나타낼 수 있는 첫 번재 주성분이 각 거시 변수에 매칭되도록 하였다. 네이버 검색 결과는 2016년부터 수집이 가능하며 2023년 중반까지의 데이터를 수집하였다.

<그림 1> 네이버 검색지수 확률(일간)



자료: 네이버 데이터랩(검색일: 2023.9.9.)을 이용해 저자가 직접 작성

그림 1은 일별로 수집된 네이버 검색지수를 나타내고 있다. 본 연구에서는 검색기간을 2016.1.1.~2023.6.30. 으로 정하고 수집하였다[3].

NDL에서는 검색량의 절대치를 공개하지 않고 해당 시기에 기록된 가장 높은 검색량을 100으로 하고 지수화하여 제공하고 있다. 또한 여러 검색어를 하나의 주제로 묶어 입력하면 해당 검색어들의 검색량을 합산한 뒤 지수로 제공하며, 한 주제에 대해 최대로 설정할 수 있는 검색어의 수는 20개로 제한되어 있다. 따라서 경제변수의 검색어가 20개를 넘어갈 경우 대표성이 있는 20개 이내의 검색어를 선정하여 지표 수집에 사용하였다.

“임금”과 관련된 검색의 경우 꾸준히 증가하여 추세를 보이고 있는 것처럼 보이지만 다른 변수의 경우 오히려 일시적으로 검색량이 폭발한 시기로 인해 검색량이 상대적으로 적은 것처럼 보인다. 하지만 이 검색지수는 해당 섹터의 절대 검색량을 최대치를 100으로 한 지수이기 때문에

3) 모든 자료의 수집 시기는 2023.9.9.이다.

다른 검색어와 수준으로서는 비교할 수 없다.

종합적으로, 8개의 거시 변수와 8개의 네이버 검색변수들이 GDP 예측 모형에 설명 변수로서 사용되었다. 일별로 수집된 검색 지수는 그림 1에서와 같이 일시적으로 검색이 폭등하였을 때로 인해 그 추이를 알기 쉽지 않다. 하지만 이 검색어들과 맞대응 되는 거시 변수들은 보통 월별로 수집되고, 본 연구의 종속변수인 GDP는 분기별로 수집되고 있기 때문에, 거시 변수들은 지수와 총량 구분에 따라 분기별로 수정하였고, 네이버 검색 변수들은 일간 데이터를 수집한 다음 이를 평균내어 월별로 전환한 후, 이를 다시 분기별 자료로 전환하였다. 추후 여러 거시 자료가 월별로도 수집 가능하기에 추후 혼합 주기 자료를 이용하여 GDP 현재 예측을 실시할 수도 있다.

그림 2는 각 항목별로 네이버 검색빈도를 분기별로 나타낸 자료이다. 검색빈도는 각 단어별로 최대 검색이 되었던 경우를 1로 하여 작성된 것이기 때문에 각 항목별로 확률이 동일하다는 것이 검색횟수가 동일하다는 것은 아니고, 상대적으로 해당 시기에 해당 검색어가 다른 어떤 때 보다 검색이 많이 되었다는 뜻이 된다. “물가” 관련한 검색어 비중에서 보면 팬데믹 이전 물가수준이 디플레이션 조짐을 보이자 관련어 검색이 급증한 것으로 나타났다. 이로 인해 상대적으로 다른 시기에는 검색 비중이 낮게 나온 것으로 보인다. 이자율 관련 검색을 나타내는 “금리” 관련해서는 디플레 가능성이 보였던 시기에 상승 후 하락하였으나, 팬데믹 이후 회복기에 금리가 상승하면서 검색 비중이 역시 높아지는 모습을 보이고 있다.

<그림 2> 네이버 검색지수 확률(월간)

자료: 네이버 데이터랩(검색일: 2023.9.9.). 일간자료를 월평균으로 변환하여 저자가 직접 작성

2. 예측 성과 비교

본 연구에서는 GDP 성장률을 대상으로 DMA와 DMS방법을 이용하여 네이버 검색 지수가 예측 모형의 성과에 얼만큼 개선시키는지를 실증적으로 분석하였다. 표 1에서 먼저 튜닝 파라미터인 λ와 α모두 0.95, 0.97 그리고 0.99로 실증 분석을 실시하였다. 앞서 언급하였듯이 0.95로 선택하면 5년 전의 관측값이 최근 관측값에 35% 정도 영향을 받는 것으로 해석할 수 있다. 그리고 네이버 검색 지수를 얼마만큼 포함할지를 다루는 ω의 선택은 0, 0.25, 0.5, 0.75 그리고 1의 경우 모두를 살펴보았다. 한편 비교 분석을 위해서 시변 모수를 포함한 DMA와 DMS외에도 같은 설명변수들을 사용한 OLS, AR(2), 그리고 임의보행 모형을 통한 전망을 실시하였다.

모형의 성과는 평균제곱예측오차(Mean Squared Forecast Error, MSFE)로 평가를 하였고, 표 2에서 기준 모형인 AR 모형에 의한 예측만이 실제 MSFE이고, 나머지 모형은 이에 대한 비율로서 표시되었다. 즉. MSFE가 1보다 크다면 AR 모형보다 더 좋지 못한 예측을 한 것이고, MSFE가 1보다 작다면 AR 모형보다 더 좋은 예측 결과를 보였다는 의미이다. i번째 모형에 대한 MSFE는 다음과 같이 정의된다.

$$MSFE_i = \sum_{t=1}^{T} y_t - \hat{y}_t^i \tag{14}$$

여기서 $\hat{y}_t^i$는 모형 i의 점 예측값이다. 또한 예측 모형의 성과에 대한 통계적 유의성을 검증하기 위해 Diebold와 Mariano(1995)의 테스트를 사용한다. 이 검증은 예측 오차와 관련된 손실 함수인 ϵ_t를 $L(\epsilon_t)$와 ϵ_t^2의 이차함수를 통해 실시한다. 예를 들어 두 예측을 비교하려면 손실 차이를 다음과 같이 정의할 수 있다.

$$d_{12,t} = L(\epsilon_{1t}) - L(\epsilon_{2t}) \tag{15}$$

이 검정에서 핵심 가정은 식 (15)의 공분산이 정상이어야 한다는 것이다. 즉 모든 t에 대해서 손실의 예상값은 어떤 상수, μ로서 $E(d_{12,t}) = \mu$로 나타낼 수 있어야 한다. 그리고 자기공분산은 시차항의 함수로, $cov(d_{12,t}, d_{12,(t-\tau)}) = \gamma(\tau)$로, 분산항은 유한하고, $0 < var(d_{12,t}) = \sigma^2 < \infty$의 조건을 만족해야 한다. 본 검정에서의 귀무가설은 두 모형의 예측 정확도가 동일하다는 것으로, $E(d_{12,t}) = 0$을 의미하고, 귀무가설하의 검정 통계량은 다음과 같이 정의된다.

$$DM = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \tag{16}$$

여기서 $\bar{d}_{12} = \frac{1}{T}\Sigma_{t=1}^{T} d_{12,t}$로 표본 평균 손실차이이고, $\hat{\sigma}_{\bar{d}_{12}}$는 $\bar{d}_{12}$의 표준편차의 추정량이다. 식 (16)의 극한 분포는 $N(0,1)$이다

참고로 대부분의 거시경제 데이터는 2000년 1월부터 사용 가능하지만, 네이버 데이터랩에서 제공하는 검색 데이터는 2016년부터 존재한다. 따라서 본 연구에서는 2016년 이후 자료들만을 대상

으로 실증 분석을 실시하였다.

<표 2> DMA와 DMS 예측 결과

		$\lambda=\alpha=0.95$		$\lambda=\alpha=0.97$		$\lambda=\alpha=0.99$	
		DMA	DMS	DMA	DMS	DMA	DMS
검색지수 이용안함	AR(2)	32.84	–	–	–	–	–
	OLS	1.57**	–	–	–	–	–
	R.W.	0.91	–	–	–	–	–
	$\omega=1$	0.79*	0.76**	0.73**	0.66**	0.69**	0.65**
검색지수 이용	$\omega=0.75$	0.84*	0.74**	0.78*	0.71**	0.73**	0.69**
	$\omega=0.5$	0.80*	0.74**	0.75**	0.71*	0.71**	0.69**
	$\omega=0.25$	0.77**	0.71**	0.73**	0.68**	0.69**	0.65**
검색지수만 이용	$\omega=0$	0.75**	0.71**	0.72**	0.68**	0.68**	0.65**
	OLS	1.82*	–	–	–	–	–

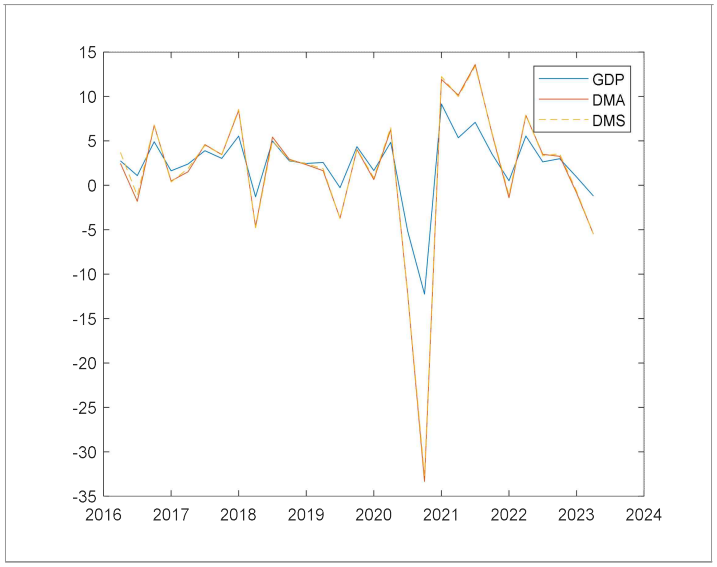
자료: 한국은행, 통계청, 네이버 데이터랩(검색일: 2023.9.9.)을 이용하여 추정한 결과에 기반하여 저자 작성.

표 2에서 검색지수를 사용하지 않은 OLS, AR(2), 임의보행 모형들의 결과들은 처음 세 행에 나타나 있다. 이 세 가지 모형들은 α나 λ값에 영향을 받지 않으므로 첫 번째 열에만 결과가 나타나 있다. AR 모형에서만 MSFE 값이 나타나 있으며, 나머지 값은 AR 모형의 MSFE에 대한 상대적 MSFE 값이다. 각 수치에서 *은 10% 수준에 DM 검증 결과 OLS 모형의 MSFE와 통계적으로 차이가 있다는 것이고, **는 5% 유의수준 하에서 역시 같은 결론을 내릴 수 있다는 뜻이다. OLS 모형은 단순하게 8개의 고려된 모든 거시 변수를 설명 변수로 사용한 모형이다. AR(2) 모형을 사용하는 것만으로도 50%정도 MSFE가 감소하는 것을 볼 수 있으며, 통계적으로도 OLS와 충분히 다른 것을 알 수 있다. 임의보행 모형은 MSFE가 오히려 AR(2)모형보다 MSFE는 낮게 나타났으나 통계적으로 유의미하게 다르지는 않은 것으로 나타났다. 검색지수를 사용하지 않은 모형 중에 $w=1$은 8개의 설명변수를 DMA와 DMS 프레임워크에 사용한 결과이다. 이에 따르면 다른 벤치마크 모형들에 비해 DMA와 DMS를 사용하는 것만으로도 예측 성과를 매우 높일 수 있었으며, 역시 AR(2) 모형과는 유의미한 차이를 보인다.

한편, 검색지수를 이용한 결과는 검색지수 결과를 사용한 비중에 따라 4가지로 나눌 수 있다. $\omega=0$은 완전히 검색지수만을 사용한 결과물이 된다. 전체적으로 튜닝 파라미터의 값과 상관없이 기존의 거시변수만을 사용한 것보다 조금 더 향상된 예측 결과를 보이고 있다. 전체적으로는 DMA보다는 DMS가 조금 더 나은 예측 결과를 보이고 있지만 큰 차이는 아니다. 검색지수를 이용하는 비율에 따라서는 완전히 검색 지수만을 이용한 경우보다는 검색지수를 보조적으로 이용한 것이 예측 결과가 조금은 더 나은 것으로 나타나지만 역시 유의미한 차이는 아닌 것으로 보인다. 참고로 마지막 행은 검색지수만을 이용하여 OLS 모형을 적용한 결과로 검색지수만을 단순하게 설명변수로 이용한 모형의 예측결과는 그리 좋지 못한 것으로 나타나 검색지수가 직접 설명변수로서의 역할은 하지 못하는 것으로 나타났다.

한편 튜닝 파라미터인 λ와 α값은 그 값이 클수록 현재 정보에 대한 가중치가 더 높다는 것을 의미하는데, 예측 성과에서는 이 값이 더 높을수록 더 좋은 것으로 나타났다.

<그림 3> GDP성장률과 DMA와 DMS 모형의 예측결과 비교



자료: 한국은행(검색일: 2023.9.9.)과 네이버 데이터랩(검색일: 2023.9.9.) 자료를 바탕으로 추정한 결과에 기반하여 저자 작성.

그림 3은 GDP의 분기별 성장률(QoQ)를 DMA와 DMS로 예측한 결과를 나타낸 것이다 ($\lambda = \alpha = 0.99$). 그림 3에서 DMA와 DMS가 큰 차이는 보이지 않아 그림 상으로는 구분하기 쉽지 않다. 보통 예측 모형들이 평균 회귀적인 성향을 보이는 데에 비해 DMA와 DMS는 변곡점들을 잘 예측하는 것으로 나타나 경기변동 시점에 더 유용하게 사용될 것으로 보인다.

결론적으로 GDP를 예측하기 위해 거시변수만을 사용하거나 AR모형을 사용하는 것 보다는 DMA나 DMS를 사용하는 것이 예측력이 개선되는 점을 확인하였다. 또한 이런 거시변수를 보조하기 위해 검색지수를 추가하는 경우에도 예측력이 거시변수만을 사용하는 DMA나 DMS보다 예측력이 개선되었다. 다만 검색지수를 사용하는 것이 이를 사용하지 않는 모형보다 통계적으로 유의미한 만큼 개선되지는 않는 것으로 보여 이에 대한 연구가 더 필요하다. 또한 거시변수를 제외하고 설명변수로 검색지수만을 사용한 모형도 예측력이 우수한 것으로 나타났는데 이는 검색지수가 보조지표가 아닌 예측을 위한 설명변수로 사용될 수 있음을 의미한다.

다만 이와 같은 경우에 다른 거시변수와 달리 각 검색지수 변수는 이론적인 함의가 담겨져 있지 않은 통계값에 불과하기 때문에 이에 대해 분석하는 것이 쉽지 않다. 그렇기에 DMA 모형에서 각 변수가 GDP 예측 모형에 기여하는 가중치를 나타내는 식(11)에서의 $q_{t|t-1,j}$를 통해 어떤 검색지수들이 GDP 예측에 많이 기여를 하는지를 파악할 수 있다. 그림 4와 그림 5는 검색지수만을 사용했을 때($\lambda = \alpha = 0.99$) 표본 외 예측을 실시하였을 때 각 검색지수들이 GDP 예측에 기여하는 바를 나타낸 것이다. 각 지수들의 변동성을 잘 나타내기 위해 모든 지표들을 독립적으로 나타냈으나 수직축의 스케일을 주의해서 봐야 한다.

8개의 지표 중 꾸준하게 높은 확률로 DMA 모형에 기여하는 변수들은 실업률과 금리지표를 나타내는 Term Spread 그리고 산업생산으로 거시변수만을 고려할 때 중요도가 높을 것으로 보이는 세 가지 변수와 매우 유사하게 선정되었다.

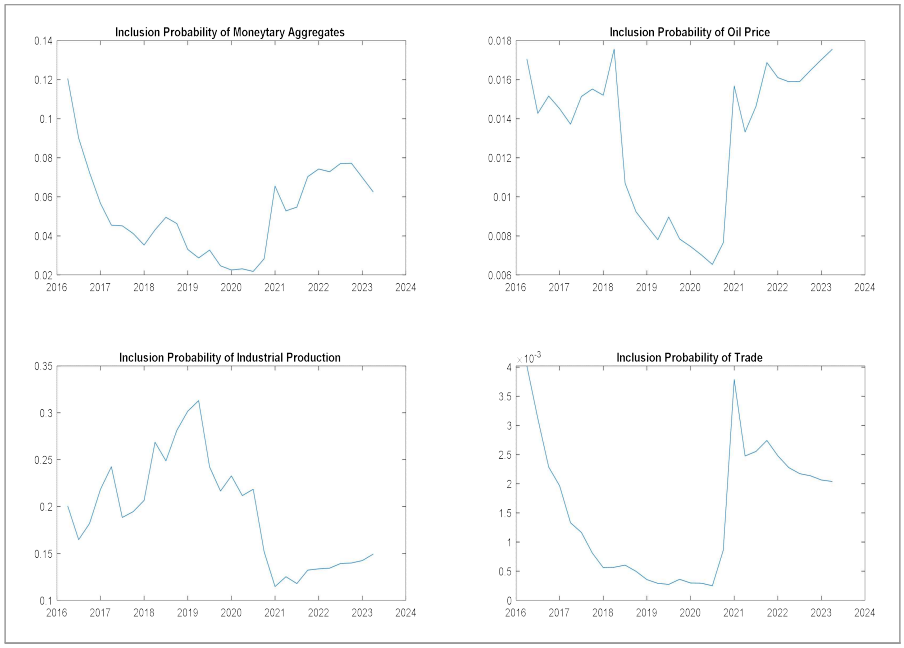
<그림 4> 검색 카테고리별 DMA모형에 포함되는 각 카테고리의 가중치

자료: 네이버 데이터랩(검색일: 2023.9.9.) 자료를 바탕으로 추정한 결과에 기반하여 저자 작성.

3. 전망 결과

검색지수를 이용한 전망모형의 실제 예측 결과를 검증하기 위하여 2022년 4분기까지의 자료를 바탕으로 모형을 training 시킨 후에 이를 다시 2023년 1~4분기의 GDP를 예측하는 표본 외 예측을 실시하였다. 검색 자료는 실시간으로 얻을 수 있어 가장 최근의 자료까지 얻을 수 있지만, 예측 주기에 따른 예측 결과를 검증하기 위하여 2022년 12월 31일까지만의 자료를 사용하여 1분기~4분기 예측을 실시하였다.

<그림 5> 표본 외 예측 결과



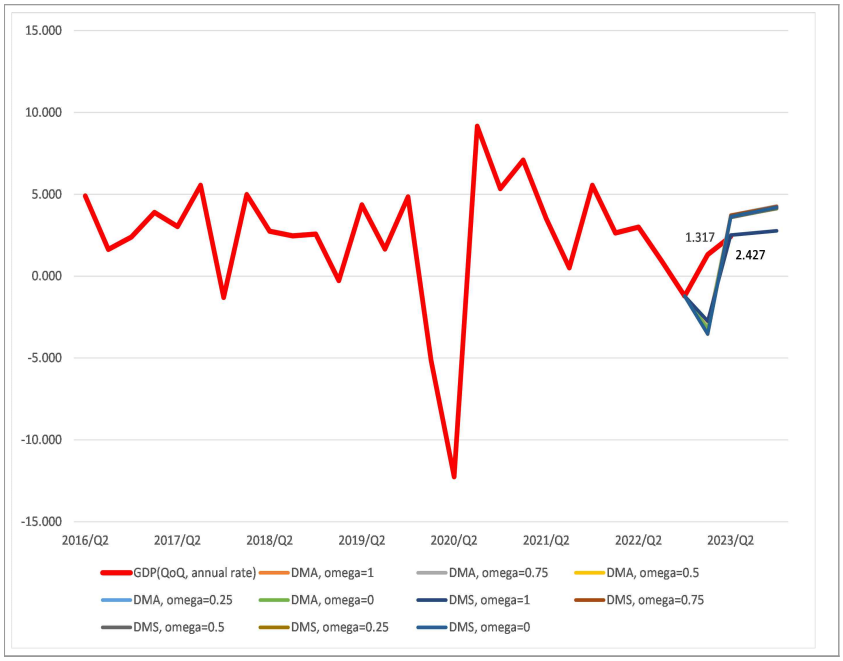
자료: 네이버 데이터랩(검색일: 2023.9.9.) 자료를 바탕으로 추정한 결과에 기반하여 저자 작성.

설명변수로 사용되는 거시변수 역시 1~4분기 후를 따로 projection하지 않고, h-step 이후의 예측을 다음과 같이 실시하였다.

$$E(y_{t+h}|y_{1:t}) = \sum_{l=1}^{J} q_{t|t-h,l} z_{t-h}^{(l)} \hat{\theta}_{t-h}^{(l)} \qquad (17)$$

즉. 2022년 4분기까지 완료된 자료를 바탕으로 2023년 1~4분기 예측을 하게 되는데, 실제로는 검색지수를 제외하고는 12월 31일 시점에서 거시 자료는 입수가 불가능하다고 할 수 있다. 가장 빨리 발표가 되는 물가지수의 경우에도 2022년 12월 자료는 영업일 기준 최소한 3일은 지나야 입수가 가능하기 때문이다. 이와 같이 실제 입수 가능한 자료를 사용하여 예측을 하기 위해서는 실시간(real-time) 데이터를 이용하여야 한다. 이를 위해서는 설명변수가 되는 GDP부터 데이터의 빈티지를 고려하는 실시간 데이터를 구축해야 하는데, 이에 대한 설명은 Kim and Swanson(2018)를 참조하길 바란다.

<그림 6> DMA, DMS의 2023년 1~4분기 예측 결과



자료: 한국은행(검색일: 2023.9.9.)과 네이버 데이터랩(검색일: 2023.9.9.) 자료를 바탕으로 추정한 결과에 기반하여 저자 작성.

그림 6은 2022년 12월말 기준 DMA, DMS를 검색지수의 가중치에 따라 2023년 1~4분기를 예측한 결과이다. 2023년 1~2분기는 이미 실현된 값이므로 그림에는 두 기간의 값이 포함되어 있다. 2022년말을 기준으로 2023년 1분기의 GDP성장률은 하락을 예상하고 있었지만 실제로는 반등한 모습을 보였다. 그에 비해 2023년 2분기에는 반등을 예상하였는데, 실제로는 2.4%에 그쳐 예상치보다는 조금 낮은 수준이었다. 한편 2023년 3~4분기는 3% 후반에서 4% 초반으로 예측을 하고 있다. 참고로 모든 GDP 성장률은 분기별 성장률을 연간 성장률로 전환한 값을 사용하였다.

Ⅳ. 시사점 및 결론

본 연구는 검색엔진의 검색 결과를 비정형 데이터로 정의하고, 이를 기존에 알려진 거시경제 변수를 예측할 수 있는 모형에서 어떤 정보를 더 추가할 수 있을지를 예측 결과를 통해 살펴보았다. 검색 결과 그 자체는 비록 정확한 경제 정보를 반영하는 것은 아니지만, 정형 데이터라고 할 수 있는 기존의 거시경제 데이터가 추적하지 못하는 여론의 흐름 혹은 소비자들의 경제심리 등을 반영할 수 있기 때문에 기존 경제 모형에 보조적인 지표로서 활용된다면 예측력 향상을 기대할 수 있을 것으로 보인다. 매우 단순한 형태의 거시경제 예측모형에서도 검색 결과 지수는 예측력을 향상시키는 결과를 가져온 것으로 이러한 기대를 확인할 수 있었다.

비정형 데이터를 예측모형에 보조적인 지표로서 활용되기 위해서는 단순히 검색 결과 지표만 반영할 것이 아니라 소비자 심리 혹은 여론 등을 조금 더 정밀하게 반영할 수 있는 텍스트 마이닝 기반의 데이터베이스를 구축하는 것이 필요할 것으로 본다. 또한 비정형 데이터 구축에 있어 타깃으로 하는 종속변수의 자료생성 프로세스에 적합하도록 알고리즘을 개발하는 것도 필요하다.

예를 들어서 Lexicon 기반의 단어사전을 구축하여 GDP 성장률을 예측하는 데 필요한 단어들을 수집하거나, 구글 등에서 검색 알고리즘으로도 사용하는 BERT와 같은 알고리즘을 이용하여 텍스트 기반의 데이터에 GDP 예측에 적합하도록 감성지수 등을 구축하는 등의 방안을 생각해볼 수 있다. 이와 같은 작업들은 의미있는 기초 자료를 구성하기 위한 노동력이 필요한 라벨링 작업과 데이터를 처리할 수 있는 컴퓨터 용량 등이 확보되어야 한다.

참고문헌

Carri`ere-Swallow, Y. and Labb´e, F. (2011). Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289.298.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Records*, 88(s1):2.9.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253.63.

Jang, I. and Choe, Y. C. (2016). Forecasting agri-food consumption using the keyword volume index from search engine data. In Agricultural and Applied Economics Association Annual Meeting, Boston, Massachusetts., number 236124.

Kim. H. H. and Swanson, N. R. (2018). Methods for Backcasting, Nowcasting and Forecasting Using Factor-MIDAS: With an Application to Korean GDP , *Journal of Forecasting*, 37(3):281-302

Koop, G. and Korobilis, D. (2012). Forecasting inflationation using dynamic model averaging. *International Economic Review*, 53:867.886.

Koop, G. and Onorante, L. (2019). Macroeconomic nowcasting using google probabilities,. Advances in Econometrics, in: Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A, volume 40, pages 17-40, Emerald Group Publishing Limited.

Lee, G. and Hwang, S. (2014). Business cycle indicator using big data: Compilation of the naver search business index. *Bank of Korea Economic Analysis*, 20(4):1.37.

Raftery, A., Karny, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52:52.66.

RiskMetrics (1996). Technical Document (Fourth Edition).

Scott, S. L. and Varian, H. R. (2013). Bayesian variable selection for nowcasting economic time series. Technical report, NBER Working Paper, No.19567.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11.30.

Vosen, S. and Schmidt, T. (2011). Forecasting private consumption: Survey-basedindicator vs. google trends,. *Journal of Forecasting*, 30(6):565.578.

West, M. and Harrison, J. (1997). Bayesian nowcasting and Dynamic Models, second edition, New York: Springer.

Yoo, D.-I. (2016). Vegetable price prediction using atypical web-search data. In Agricultural and Applied Economics Association, Annual Meeting 2016, Presentation No. 236211.

부록. 거시경제 변수별 검색어 목록

영문 검색어는 Koop and Onorante(2019)에서 Google을 기준으로 사용된 검색어이고, 이에 대응되는 한국어를 해당 분야별로 정리하였다. 해당 한국어 검색어를 바탕으로 NDL을 통해 네이버 검색지수를 생성하였다.

<부표> 거시경제 변수별 검색어 목록

(단위: 내용)

주제	검색어(한국어)	검색어(영어)
Wage Inflation	임금, 급여, 국내 임금수준, 임금 물가상승, 월급 인상, 급여 수준, 월급 물가상승률, 월급 비교, 시간당 임금 계산기, 월급 계산기, 시간당 임금, 연봉 계산기, 평균 연봉, 산재 보상, 급여 물가상승, 월급 세금 계산, 급여세 계산기, 임금 인플레이션, 실질 임금	wages, wages calculator, wage, wage inflation, us wages, salary, salary tax calculator, salary raise, salary grade, salary comparison, salary calculator hourly, salaries, real wages, paycheck calculator, compensation, calculate salary
Term Spread	이자율, 금리, 국내금리, 한국은행, 실질금리, 우대 금리, 우대 대출금리, 주담대 금리, 기준금리, 이자율 물가상승, 금리 인플레이션, 금리 추세, 금리 인상, 금리 인하, 이자율 전망, 이자율 계산기, 이자율 계산, 주담대 이자율, 한국은행 금리, 이자율 예측	the fed, real interest rate, prime rate, prime interest rate, mortgage rate, mortgage interest rates, interest rates, interest rates inflation, interest rate, interest rate predictions, interest rate inflation, interest rate forecast, interest rate fed, interest rate drop, interest rate cuts, interest rate cut, interest rate claculator, feds interest rate, federal interest rate, fed, fed rates, fed rate, fed rate cut, fed interest rates, fed interest rate
Unemployment Rate	한국 실업, 국내 실업, 한국 실업률, 국내 실업률, 실업, 실직, 실업 통계, 실업자 통계, 실업률, 실업 급여 신청, 실업수당 신청, 실업 보험, 실업 급여, 실업 수당, 실업 연장, 전국 실업, 경기침체 실업률, 국가보조금, 국가지원금, 고용 증가	us unemployment, us unemployment rate, unemployment, unemployment statistics, unemployment rates, unemployment rate, unemployment insurance, unemployment great depression, unemployment extension, unemployment depression, unemployment checks, unemployment check, unemployment benefits, subsidies, job growth, federal unemployment, depression unemployment rate
Inflation	인플레이션, 물가, 인플레이션율, 물가상승률, 디플레이션, 국내 물가, 물가상승률 그래프, 물가 전망, 인플레이션 전망, 인플레이션 뜻, 하이퍼인플레이션, 경제	what is inflation, what is deflation, us inflation rates, us inflation rate, united states inflation, u.s. inflation, real inflation, rate of inflation, national inflation, inflation, infaltion usa, infaltion rates, inflation rate, inflation in us, infaltion forecast, inflation deflation, inflation

	인플레이션, 인플레이션 뜻, cpi, 소비자 물가 지수, 실질 인플레이션, cpi 지수, 부채 디플레이션, 인플레이션이란?, 디플레이션이란?	definition, historical inflation, high inflation, deflation, cpi, cpi index, consumer price index
Money Supply	화폐, 통화, 화폐 수요, 화폐 디플레이션, 화폐가치 물가하락, 통화가치 물가하락, 통화정책, 한국은행 통화정책, 금융통화위원회	money, money deflation, monetary policy, monetary deflation
Oil Price Inflation	석유 생산, 원유 생산, 석유 생산량, 원유 생산량, 유가, 석유 가격, 휘발유 가격, 경유 가격, 경유가, 휘발유가, 에너지 생산, 전기 생산, 에너지 가격, 에너지 요금, 전기 요금, 전기 이용료, 디젤 가격, 디젤가, LPG가격	oil production, oil prices, oil price, gasoline price, gas price, energy production, energy price, electricity price, diesel price
Commodity Price Inflation	철강가격, 철값, 철가격, 철강가, 식품 가격, 식음료가, 식품가, 식음료가격, 구리 가격, 구리 값	steel price, food price, copper price
Financial Conditions Index	주식보상, 스톡옵션, 투자은행, IB, 성장주, 성장주식, 가치주, 가치주식, 골드만삭스, Goldman Sachs	stock compensation, investment banking, growth equity, goldman sachs, equity compensation
Industrial Production	경제성장률, 국내총생산, GDP, 대공황, 공황, 중소기업, 매출, 성장 비즈니스, 경제성장, 경기 순환, 경제 위기, 생산 일자리, 제조업 일자리, 생산 기업, 생산 회사, 경기 침체, 침체, 불경기, 불황	us gdp growth, the great depression, small business growth, sales growth, recession, growth, great depression, gdp growth, economic growth, cycle, crisis, business growth, business cycle, production company, production companies
Trade	수출, 수입, 경상수지, 무역수지, 경상수지 흑자, 경상수지 적자, 무역수지 흑자, 무역수지 적자, 자본수지, 자본수지 적자, 자본수지 흑자, 환율, 달러환율, 달러화, 대미무역, 무역량, WTO, 관세, FTA, 보호무역	-