Оценка использования Google Trends для прогнозирования развития кредитования

Edwige Burdeau, Banque de France, Paris, France – edwige.burdeau@banque-france.fr

Etienne Kintzler, Banque de France, Paris, France – etienne.kintzler@banque-france.fr

Перевод: *Статкомитет СНГ*

Аннотация. Все больше литературы по прогнозированию посвящено полезности данных о поисковых запросах Google. Данные о частоте поиска терминов в Google за определенный период времени доступны почти в реальном времени через web-приложение *Google Trends*. Кроме того, все больше внимания привлекает инструмент *Google Correlate*, который показывает тенденции *Google Trends*, наиболее сильно коррелирующие с рядами, предоставленными пользователями. Для каждого рассматриваемого ряда, имеется много потенциальных прогностических параметров, которые могут использоваться для целей прогнозирования. В этой связи была проведена оценка дополнительной ценности инструментов *Google Trends* и *Google Correlate* для прогнозирования потоков кредитов с целью покупки жилья во Франции. Поскольку взаимосвязь между *Google Trends* и рассматриваемыми переменными необязательно носит линейный характер, прогностическая сила таких показателей может быть улучшена при помощи использования методов машинного обучения, особенно нелинейных. Тестируются разные популярные методы машинного обучения: линейные методы, такие как метод *LASSO (Shrinkage and Selection Operator) и байесовские методы машинного обучения* (BMA), а также нелинейные методы, такие как модели усиления (Boosting models) и *Метод опорных векторов с нелинейными ядрами* (*Support Vector Machines* (SVM) with non-linear kernels). Прогностическая сила каждой модели оценивается с использованием ошибок прогнозирования за пределами выборки и сравнивается с базовой моделью, которая не включает показатели *Google Trends*. Мы обнаружили, что полезность *Google Trends* для целей прогнозирования за несколько месяцев заранее доказана и, до некоторой степени, не зависит от выбранной модели. Кроме того, в некоторых случаях нелинейные модели демонстрируют более высокую прогностическую силу, чем линейные.

Ключевые слова: прогнозирование; Google Trends; кредитные показатели; модели выбора переменных.

Мониторинг изменений в потоках кредитов в реальную экономику имеет большое значение для центрального банка. Кредитный канал является преимущественным способом финансирования экономики для центрального банка: центральный банк может корректировать свою денежно-кредитную политику путем снижения ключевых процентных ставок. Однако более низкая процентная ставка не отражается немедленно в ставках процента по кредитам, поскольку экономическим агентам требуется несколько месяцев для получения кредита, например кредита для покупки жилья. Во Франции обычно требуется три месяца, которые даются покупателю для того, чтобы определить план финансирования после согласования предварительного соглашения о покупке/продаже. В этих условиях изменение ключевых ставок оказывает влияние на кредитные потоки через несколько месяцев после принятия решения об изменении монетарной политики. Такие показатели, как *Google Trends* становятся доступными в квазиреальном режиме и могут предоставить раннюю информацию об ожидаемом спросе на кредиты, поскольку покупатели жилья используют интернет для планирования своих покупок. Основная цель данной статьи состоит в том, чтобы оценить, является ли приложение *Google Trends* ценным инструментом прогнозирования потоков кредитов для покупки жилья. На практике использование *Google Trends* вызывает некоторые вопросы. Во многих исследованиях не описано то, как выбираются «лучшие» показатели *Google Trends*. Более того, не очевидно, что один или два термина могут быть лучшими прогностическими факторами для временного ряда, и даже если это так, то может оказаться, что эти термины будут отсутствовать в будущем, так как привычки web-пользователей изменятся. В этой работе мы предлагаем более надежный способ выбора показателей из *Google Trends,* а именно – опираясь на *Google Correlate*. Кроме того, мы тестируем прогностическую силу *Google Trends* при помощи нескольких моделей и, особенно, моделей выбора переменных, для того, чтобы оценить, зависят ли наши результаты от модели и улучшает ли результаты рассмотрение нелинейных воздействий. Хотя модели выбора переменных не так важны, когда набор переменных является разумным, эти модели позволяют правильно определить экономный набор релевантных показателей из большого набора данных. Наконец, мы проверяем прогностическую способность наших показателей *Google Trends* как по чистым потокам кредитов, так и по потокам новых кредитных контрактов на покупку жилья, поскольку соглашения об обратном выкупе могут изменять возможности *Google Trends* для корректного прогнозирования чистых потоков кредитов для покупки жилья. На основе проведения разных экспериментов, оказывается, что использование *Google Trends* может помочь в прогнозировании кредитных потоков на несколько месяцев вперед, и это свойство, похоже, не зависит от выбора модели. Наконец, наш первый эксперимент с

нелинейной моделью является убедительным, и в будущем дальнейшие усилия должны быть посвящены изучению этого вопроса, чтобы правильно идентифицировать нелинейные закономерности в этих временных рядах.

1. Литература о Google Trends

При помощи приложений *Google Trends* и *Google Correlate*, Google дает каждому возможность анализировать тенденции в размещении поисковых запросов в интернете в квазиреальном времени. Для большинства популярных терминов Google Trends показывает недельную или месячную динамику количества запросов, нормализованных по общему количеству запросов в рамках географической области или временного интервала. История расчета индексов запросов начинается с января 2004г. Google подчеркивает в своем документе, посвященном Google Correlate (2011), что в то время как выбор интересных трендов Google не представляет труда, выбор правильного набора индикаторов далеко не тривиален. В связи с этим Google разработал приложение *Google Correlate,* которое «позволяет автоматически выбирать запросы из миллионов кандидатов» для любой временной схемы. На практике *Google Correlate* можно использовать двояким образом. Пользователь может загрузить недельные или месячные ряды по своему выбору; тогда приложение выдает наиболее тесно коррелирующие тренды *Google Trends*. С другой стороны, пользователь может указать до ста терминов запросов Google, сильно коррелирующих с предопределенным индивидуальным запросом. В этом случае *Google Correlate* дает возможность определить семейство терминов близкое к тому, что получают методами обработки естественного языка.

Со времени первой публикации показатели *Google Trends* стали хорошо известны благодаря их свойствам прогнозирования на ближайшее время, а иногда и на более отдаленный период. В предварительной работе Choi & Varian (2009) подчеркивалось, что некоторые краткосрочные прогнозы могут быть улучшены при помощи использования показателей *Google Trends,* например, такие как прогнозы на ближайшее время продаж автомобилей, розничных продаж или начала строительства частного жилья. Для жилищного сектора в работах Wu & Brynjolfsson (2015) была показана практическая польза предопределенных категорий по недвижимости в *Google Trends* для диагностирования и прогнозирования продаж домов и цен на жилые дома в США. Askitas (2015) также использовал web-запросы, помещенные в категорию «Недвижимость» в *Google Trends,* для краткосрочного прогнозирования цен на жилые дома в США, а Chauvet, Gabriel & Lutz (2016) построили индекс риска дефолта по ипотеке для предвидения показателей просрочки платежей по ипотеке. Наконец, Coble & Pincheira (2017) показали, что приложение *Google Trends* может также

помочь в прогнозировании количества разрешений на строительство в США. Мы внесли свой вклад в это направление в литературе своими прогнозами потоков кредитов для покупки домов. *Google Trends* может обеспечить раннюю информацию в следующей конкретной области: будущие покупатели используют поиск в интернете для оценки стоимости кредита, поиска потенциального жилья или оценки предложений банков. Кроме того, центральный банк Франции собирает раннюю информацию на стороне банков, агрегированную в Обзоре банковского кредитования (*Bank Lending Survey),* однако на стороне домашних хозяйств имеется немного качественных показателей о кредитовании покупки жилья.

Кроме того, хотя большая часть литературы была посвящена простым моделям, сводящим использование *Google Trends* к получению дополнительных объясняющих переменных в авторегрессионных моделях, некоторые работы обогатили литературу. Для краткосрочного прогнозирования частного потребления в Германии Schmidt & Vosen (2011) сократили размерность начального набора экзогенных переменных, взяв главные компоненты из разных категорий *Google Trends*. Scott & Varian (2012) отобрали большой набор показателей *Google Correlate* в качестве входных данных для байесовской модели выбора переменных. Koop & Onorante (2013) подчеркивали возможности показателей *Google Trends* в определении поворотных точек, обращая внимание на то, что прогностическое свойство *Google Trends* не обязательно линейно. В соответствии с этой литературой мы работали над многими показателями Google Trends, выбранными Google Correlate, и суммировали их в главные компоненты. Кроме того, мы используем модели выбора переменных для ограничения шума нерелевантных показателей и тестируем нелинейный метод.

2. Описание набора моделей

Для прогнозирования потоков кредитов для покупки жилья мы решили протестировать недавно разработанные методы машинного обучения, которые могут обращаться с большим набором объясняющих переменных. Действительно, некоторые статьи в литературе по прогнозированию, такие как работы Bai & Ng (2008, 2009) и Kim & Swanson (2016), подчеркивают преимущества использования моделей из области машинного обучения.

В этой области, модели LASSO и модели эластичных сетей, предложенные Tibshirani (1996) и Zou & Hastie (2005) являются хорошо известными линейными моделями, в которых добавляется ограниченный член к функции потерь для снижения размерности параметров модели, особенно для нерелевантных переменных. LASSO представляет собой особый случай, где параметры могут быть сжаты до нуля. Сжимаемый параметр, отражающий уровень серьезности ограничения, должен быть калиброван, обычно при помощи кросс-валидации.

Хотя первоначально эти методы усиления и особенно компонентный L2 метод усиления, предложенный Buhlmann & Yu (2003) и Buhlmann & Hothorn (2007), использовались для целей классификации, они могут использоваться для регрессионных задач. Эти модели оцениваются рекурсивно, на каждом шаге оценка увеличивается на элементарную функцию одной переменной, называющуюся обучаемой функцией (распознавателем) (*base learner),* которая минимизирует функцию потерь. На практике количество шагов должно быть также калибровано при помощи процедуры кросс-валидации. Обучаемая функция может быть линейным или сплайн компонентом. В нашем случае проверяется каждый тип обучаемой функции, но только линейный компонент дает интересные результаты. Усреднение байесовских моделей (BMA), объясненное в работе Raftery & al. (1997), оценивает линейные модели всех комбинаций первоначального набора переменных или выборки, если начальное количество переменных слишком велико. Окончательная модель представляет собой взвешенную среднюю всех этих моделей. Вес каждой модели происходит из апостериорной вероятности, полученной на основе теоремы Байеса. Для этой цели для каждой модели должна быть определена априорная вероятность, которая отражает то, насколько хорошей модель считается *априорно*. В нашем случае были протестированы разные распределения априорной вероятности: равномерное, случайное или фиксированное с использованием кросс-валидации. Лучшие результаты были получены для равномерного распределения. Наконец, мы расширили данную работу, протестировав метод опорных векторов (SVM). Впервые определенный Vapnik et al. (1992) для решения проблем классификации, метод опорных векторов был адаптирован для решения регрессионных задач. Обычно SVM рассматривается как непараметрический метод, и спецификация модели опирается на функции ядра, установленные на конечном числе точек. В этом исследовании мы использовали полиномиальное ядро. В отличие от предыдущих моделей, эта модель не является моделью выбора переменных, но может быть откалибрована для того, чтобы избежать переподгонки (overfitting). Эта модель представляет собой попытку посмотреть, следует ли далее изучать нелинейные модели.

3. Набор данных

Интересующая нас данные – ряды потоков кредитов для покупки жилья и новые контракты кредитования для покупки жилья – извлечены из статистического веб-сайта Банка Франции (Webstat). Дальнейшей трансформации данных не требуется: месячные ряды, имеющиеся в открытом доступе, являются сезонно сглаженными. Для *Google Trends* ряды *Google Correlate* используются двояким образом. Первый набор показателей получают, отбирая

сначала 100 терминов, которые наиболее сильно коррелируют с терминами «кредит на покупку жилья» и «предоставление ссуды на покупку жилья» (на французском языке). Полученные термины четко связаны с отличительными особенностями покупки жилья, то есть такими терминами, как процентные ставки по кредитам, кредитное страхование или названия кредитных организаций. Затем мы извлекаем пять первых главных компонентов, рассчитанных для этого набора переменных, и выделяем нестационарные. Мы пытались включить 10 главных компонентов, но дополнительные главные компоненты не продемонстрировали интересных связей. Если первые главные компоненты имеют тенденцию к росту, вполне объяснимо подхватывая общую тенденцию, то следующие компоненты, и особенно второй, отражают идиосинктратические тенденции жилищного сектора. Этот второй компонент, по-видимому, предвосхищает изменения кредитных потоков для покупки жилья, и особенно количество новых контрактов (рисунок 1).

Рисунок 1: Изменения чистых потоков кредитов на покупку жилья, новых контрактов на кредитование покупки жилья и второго главного компонента, полученного из 100 коррелятов, связанных с запросами о кредитовании покупки жилья. Каждый ряд был стандартизован и сглажен по трехмесячным периодам для наглядности.

Источники: Банк Франции, Google Trends

Приложение *Google Correlate* также используется для определения временных рядов, коррелирующих с загруженными рядами внешних данных о кредитах для покупки жилья. В этом исследовании только непогашенные суммы и годовые темпы роста кредитов для покупки жилья коррелируют достаточно хорошо с некоторыми трендами в *Google Trends* для получения интересных результатов. Кроме того в наборе, выбранном при помощи *Google Correlate*, релевантными являются только несколько индикаторов. Вы выбрали 12 индексов *Google Trends,* связанных с терминами, содержащими названия французских кредитных организаций.

Индексы *Google Trends,* используемые в наших моделях, сезонно сглажены и учитывают первые разности нестационарных показателей с порогом 10%. Окончательный набор экзогенных переменных содержит как главные компоненты первого набора, так и индексы Google Trends из второго набора.

4. **Основные результаты**

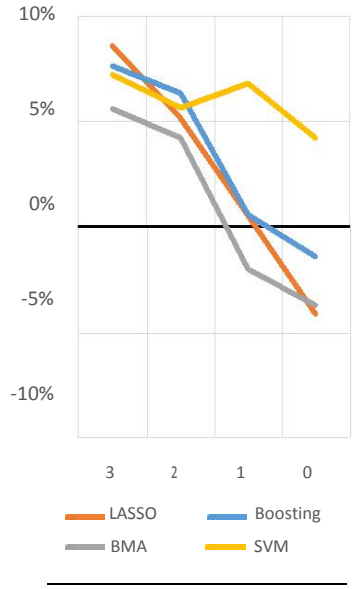
Цель нашего исследования состояла в том, чтобы оценить прогностические возможности *Google Trends*. Мы оцениваем для двух рассматриваемых переменных упомянутые выше модели: модель LASSO, эластичные сети, L_2 подход к усилению, подход BMA и модель SVM с полиномиальным ядром. Для каждого рассматриваемого месячного ряда данных, мы получаем четыре набора оценок в зависимости от количества месяцев (*h*) (от 3 до 0) до прогноза. Для каждого набора оценок мы прогнозируем значение рассматриваемой переменной в месяц *m* на основе информации, имеющейся на *h* месяцев раньше, наблюдаемых значений зависимой переменной в месяцы *m-h-1, m-h-2, m-h-3* и месячные индикаторы *Google Trends* за месяц *m-h*. Для оценки прогностической возможности наших моделей по всем моделям были рассчитаны месячные прогнозы с использованием имевшихся данных с января 2012 г. по декабрь 2016 г. Эти прогнозы мы получали рекурсивно: для каждого месяца внутри этого временного диапазона мы повторно оценивали и калибровали каждую модель, рассматривая набор данных с февраля 2004 г. по дату последнего наблюдения и прогнозировали следующее наблюдение. На основании этих прогнозов оценивается прогностическая возможность каждой модели с использованием среднеквадратической ошибки по прогнозному периоду. Чтобы сравнить эти результаты, мы также оцениваем прогнозы вне выборки в том же диапазоне периодов только с запаздывающими значениями интересующей переменной, имеющимися на момент предвидения; эта модель обозначается аббревиатурой «AR».

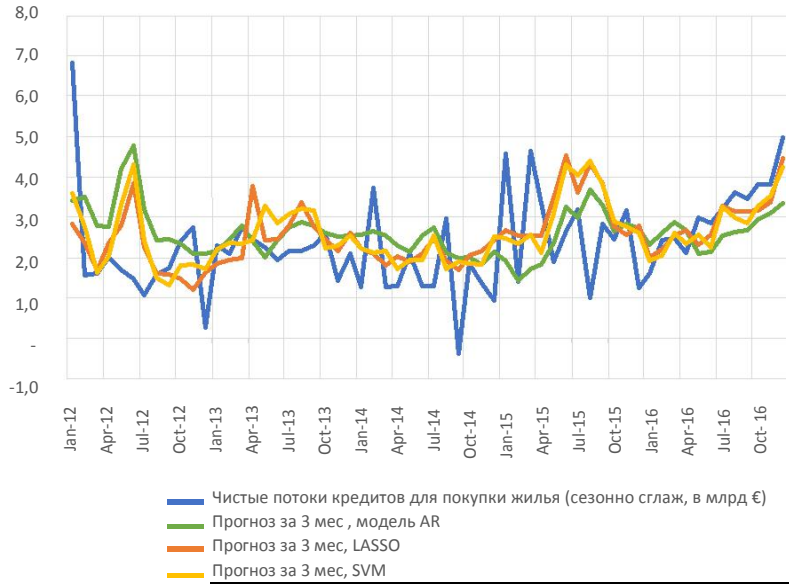
Результаты для чистых потоков кредитов для покупки жилья и новых контрактов по кредитованию покупки нового жилья показаны на рис. 2 и 3. В этом исследовании для новых контрактов мы не использовали первые разности, хотя мы не можем исключить гипотезы о единичном корне при рассмотрении кредитных изменений в 2016 г. Все же мы решили показать этот результат, поскольку это может проиллюстрировать более высокую прогностическую способность Google Trends для новых контрактов, особенно в 2013 году, а не для чистых кредитных потоков. Действительно, чистые кредитные потоки не включают в себя реструктуризованные кредиты, в то время как некоторые запросы Google явно связаны с контрактами такого типа. Из этих результатов видно, что Google Trends имеет ограниченную прогностическую силу для краткосрочного периода для обеих рассматриваемых переменных;

улучшения среднеквадратической ошибки невелики, обычно меньше 5% (рис. 2a и 3a). Но представляется, что индикаторы *Google Trends* обладают прогностической силой для более продолжительных периодов для обоих рядов, особенно за несколько месяцев вперед. В этом случае улучшения среднеквадратической ошибки довольно значительны для каждой модели и каждого ряда, от 5% до более 20% в одном случае (рис 2a и 3a). Кроме того, представляется, что индикаторы *Google Trends* дают существенную информацию для лучшего отслеживания поворотных точек цикла, особенно в 2012 г. в первой половине 2015 г. и в 2016 г. (Рис. 2b и 3b). Наилучшим прогностическим фактором из *Google Trends,* выявленным по всем моделям, действительно, является второй компонент из 100 коррелятов, показанный на рис. 1, параметр при этой переменной является положительным для всех моделей и имеет тот же порядок величины, что и член авторегрессии. Наконец, рассмотрение нелинейных характеристик с помощью SVM может значительно улучшить результаты прогнозирования.

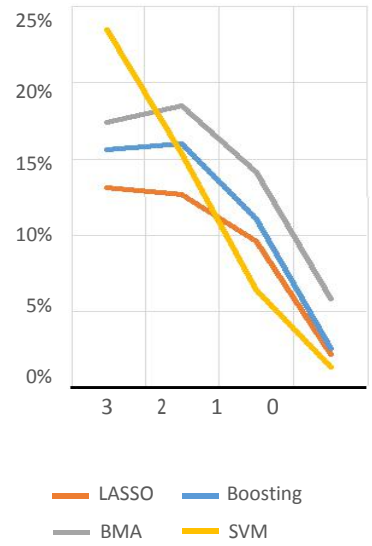
Наши эксперименты подтверждают, что использование *Google Trends* и *Google Correlate* для целей прогнозирования потоков кредитов для покупки жилья могут существенно улучшить среднесрочные прогнозы. Благодаря простому подходу использования *Google Correlate* для определения семейства терминов и анализа главных компонентов для суммирования информации, мы получили надежный индикатор будущих изменений кредитов для покупки жилья. Эти результаты говорят о том, что необходимо дальнейшее исследование некоторых аспектов. Во-первых, *Google Trends* может быть более ценным инструментом для отслеживания циклов средней частоты, чем высокой частоты; анализ *Google Trends* в рамках частотного подхода может открыть новые горизонты. Во-вторых, индексы *Google Trends* могут быть более интересны для уровней, чем для первых разностей; следует отдавать предпочтение моделям, справляющимся с нестационарностью. Наконец, наш первый эксперимент с нелинейными спецификациями дал удовлетворительные результаты, которые свидетельствуют в пользу использования других нелинейных методов.

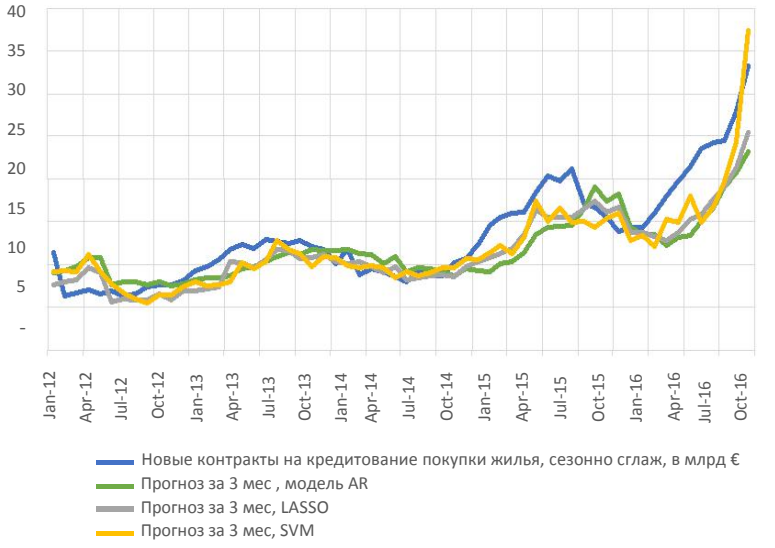
Рисунки 2a и 2b: Чистые потоки кредитов на покупку жилья – Улучшение среднеквадратической ошибки по сравнению с моделью AR от количества месяцев до прогноза, для каждого типа модели (только лучшие модели каждого типа представлены) и Сравнение результатов для рассматриваемой переменной и прогнозов за 3 месяца по разным типам моделей (справа)





Рисунки 3a и 3b: Новые контракты на кредитование покупки жилья – Улучшение среднеквадратической ошибки по сравнению с моделью AR от количества месяцев до прогноза, для каждого типа модели (только лучшие модели каждого типа представлены) и Сравнение результатов для рассматриваемой переменной и прогнозов за 3 месяца по разным типам моделей (справа).





154

Литература

Askitas, N. (2015). Trend-Spotting in the Housing Market. IZA Discussion Paper No. 9427.

Bai, J., &Ng, S. (2008). Forecasting economic time series using targeted predictors. Journal of Econometrics, Elsevier, vol. 146(2), p. 304-317.

Bai, J., &Ng, S. (2009). Boosting diffusion indices. Journal of Applied Econometrics, John Wiley & Sons, Ltd., vol. 24(4), p. 607-629.

Boser, B. E., Guyon I. M., &Vapnik V. N. (1992). A training algorithm for optimal margin classifiers.

Proc. 5th Annu. Workshop on Comput. Learning Theory, ACM Press, p. 144-152.

Buhlmann, P., &Yu, B. (2003). Boosting with the L2 Loss: Regression and Classification, Journal of the American Statistical Association, 98, issue, p. 324-339.

Buhlmann, P., &Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting.

Statistical Science, Vol. 22, No. 4, p. 477-505.

Chauvet, M., Gabriel, S., &Lutz, C. (2016). Mortgage default risk: New evidence from internet search queries. Journal of Urban Economics, 96, November, p. 91–111.

Choi, H., &Varian, H. (2009). Predicting the Present with Google Trends, Technical report, Google. Coble, D., &Pincheira, P. (2017). Nowcasting Building Permits with Google Trends. MPRA Paper. Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H. & Kumar, S (2011). Google Correlate Whitepaper

Kim, H. H., &Swanson, N. R. (2016). Mining Big Data Using Parsimonious Factor, Machine Learning, Variable Selection and Shrinkage Methods. International Journal of Forecasting.

Koop, G., &Onorante, L. (2013). Macroeconomic nowcasting using Google probabilities. Mimeo

Raftery, A. E., Madigan, D. &Hoeting J. A. (1997). Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, Vol. 92, n° 437, p. 179-191.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society Series B, 73, issue 3, p. 273-282.

Scott, S., &Varian H. (2012). Predicting the present with Bayesian structural time series. Tech. Google. Vosen, S., &Schmidt T. (2011). Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. Journal of Forecasting, Vol. 30, n° 6, p. 565-578.

Wu, L., &Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. Economic Analysis of the Digital Economy, p. 89–118.

Zou, H., &Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67, issue 2, p. 301-320.