5/22
ПРЕПРИНТЫ

ПРОСТРАНСТВЕННОЕ ПРОСТРАНСТВЕННОЕ РАЗВИТИЕ. РЕГИОНАЛЬНАЯ РАЗВИТИЕ. РЕГИОНАЛЬНАЯ И ГОРОДСКАЯ ЭКОНОМИКА И ГОРОДСКАЯ ЭКОНОМИКА SPATIAL DEVELOPMENT SPATIAL DEVELOPMENT REGIONAL AND URBAN ECONOMY REGIONAL AND URBAN ECONOMY

К. В. Ростислав, Ю. Ю. Пономарев
Д. М. Радченко

РАЗРАБОТКА ПОДХОДА К ОЦЕНКЕ ОТНОСИТЕЛЬНОЙ СИЛЫ МЕХАНИЗМОВ АГЛОМЕРАЦИОННЫХ ЭФФЕКТОВ В РОССИИ НА ОСНОВЕ МИКРОДАННЫХ О РОССИЙСКИХ ПРОИЗВОДИТЕЛЯХ И МУНИЦИПАЛЬНЫХ ОБРАЗОВАНИЯХ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

К. В. Ростислав, Ю. Ю. Пономарев, Д. М. Радченко

РАЗРАБОТКА ПОДХОДА К ОЦЕНКЕ ОТНОСИТЕЛЬНОЙ СИЛЫ МЕХАНИЗМОВ АГЛОМЕРАЦИОННЫХ ЭФФЕКТОВ В РОССИИ НА ОСНОВЕ МИКРОДАННЫХ О РОССИЙСКИХ ПРОИЗВОДИТЕЛЯХ И МУНИЦИПАЛЬНЫХ ОБРАЗОВАНИЯХ

Москва 2022

Авторы:
Пономарев Юрий Юрьевич кандидат экономических наук, Руководитель Центра пространственной экономики ИПЭИ РАНХиГС
Радченко Дарья Максимовна Лаборатория инфраструктурных и пространственных исследований ИПЭИ РАНХиГС, младший научный сотрудник
Ростислав Кирилл Владимирович Лаборатория инфраструктурных и пространственных исследований ИПЭИ РАНХиГС, младший научный сотрудник

Аннотация
Развитие агломераций в России — приоритет пространственной политики. Для усиления агломерационных эффектов и ускорения роста российской экономики нужно понимать механизмы агломерационных эффектов. Для сравнения силы агломерационных эффектов по Маршаллу с помощью подхода Эллисона — Глейзера — Керра была измерена степень сосредоточения российских отраслей по данным обо всех без исключения организациях на 1 января 2020 г. Оценки показывают, что в России пары отраслей обычно рассредоточены относительно друг друга: у большей части отраслей сосредоточение существенно ниже, чем можно было ожидать, исходя из общего размещения этих производств. В среднем из трех внешних выгод сосредоточения по Маршаллу в России важнее всего большой рынок труда. Близость к поставщикам/покупателям, их разнообразие меньше всего связано с размещением отраслей в тех же районах. На примере Калининградской области показано, что вне зависимости от способа отбора организаций для сравнения признаков усечения распределения нет. Для проверки этого вывода мы использовали различные способы территориальной группировки предприятий. В частности, были оценены с помощью метода DBSCAN границы кластеров (агломераций) предприятий. Полученные оценки связи сосредоточения с различными источниками его внешних выгод подкрепляют те меры государственной политики, которые направлены на поощрение развития крупных городских агломераций с большим и постоянным рынком квалифицированного рабочего труда. При формировании особенно плотных кластеров целесообразно устанавливать для территорий с особым режимом предпринимательства требования к виду деятельности, которые бы согласовывались с оценками интенсивности возможного обмена знаниями между отраслями.

The development of agglomerations in Russia is a priority of spatial policy. To enhance agglomeration effects and accelerate the growth of the Russian economy it is necessary to understand the mechanisms of agglomeration effects. To compare the strength of Marshall agglomeration effects using the Ellison-Glaser-Kerr approach, the degree of concentration of Russian industries was measured using data on all organizations without exception as of January 1, 2020. The estimates show that pairs of industries in Russia tend to be dispersed relative to each other: most industries have significantly lower concentration than would be expected based on the overall location of these industries. On average, of the three external benefits of concentration according to Marshall, Russia's large labor market is the most important. Proximity to suppliers/buyers, their diversity is least related to the placement of industries in the same areas. The example of Kaliningrad region shows that regardless of the method of selection of organizations for comparison, there is no truncation of the distribution traits. Although the choice of the geographical unit of observation determines the estimation of the strength or even direction of the net agglomeration effects, the general conclusion about the lack of selection of enterprises, which we could take for the benefit of concentration, was unchanged. To verify this conclusion, we used various methods of territorial grouping of enterprises and the boundaries of clusters (agglomerations) of enterprises were estimated using the DBSCAN method. The resulting estimates of the relationship of concentration to various

sources of its external benefits support those public policies that seek to encourage the development of large urban agglomerations with large and constant markets for skilled labor. When forming particularly dense clusters, it is advisable to set activity requirements for areas with a special entrepreneurial regime, which would be consistent with estimates of the intensity of possible knowledge exchange between industries.
Ключевые слова: агломерации, агломерационные эффекты, механизмы, делимитация границ, машинное обучение
Keywords: agglomerations, agglomeration effects, mechanisms, boundary delimitation, machine learning
JEL: R1, C02

Содержание

1. Виды агломерационных эффектов

Понятие агломерационных эффектов весьма расплывчатое. В самом широком смысле они означают все те экономические явления, которые возникают из-за сосредоточения производства, скопления производителей в одном месте. В более узком смысле агломерационными эффектами называют только внешние выгоды и издержки от такого сосредоточения.

Широту и смысл агломерационных эффектов легче всего представить с помощью различных способов их классификации. Такие классификации важны не только потому, что проясняют существо агломерационных эффектов, но также потому, что, в сущности, определяют способ их изучения.

Прежде всего различают положительные и отрицательные агломерационные эффекты. Так как сосредоточение производства в городах очевидно, особенное внимание привлекают выгоды сосредоточения.

Наиболее старые, а потому и влиятельные в литературе две классификации внешних выгод сосредоточения производства: одна восходит к работе Альфреда Маршалла «Принципы политической экономии» (1890), а другая — к работе Бертиля Улин (Олина) 1935 г. и Эдгара Мэлоуна Гувера 1937 г. Разделение агломерационных эффектов по Маршаллу во главу угла ставит то, перемещение чего сосредоточение производства удешевляет, а разделение (условно) Улина — Гувера — то, от скопления каких (похожих или нет) производителей возникают выгоды.

По Маршаллу различают выгоды:

— для/от постоянного рынка квалифицированного труда (работнику легче найти оплачиваемое место, а производителю — найти нужного специалиста);
— для/от большого местного рынка промежуточных товаров (ближе поставщики и покупатели, выше их разнообразие, а значит устойчивость поставок);
— от перетоков знаний (часто этот вид выгод иллюстрируют цитатой Маршалла: «[в промышленных районах] тайны профессии перестают быть тайнами, они как бы пронизывают всю атмосферу») [1, p. 352].

По Улину — Гуверу различают выгоды от сосредоточения:

— на уровне отдельного производителя (внутренний эффект масштаба);
— внутриотраслевые выгоды (от сосредоточения похожих предприятий, от местной специализации, также т. н. эффекты локализации);
— межотраслевые выгод (от сосредоточения производства и населения вообще, т. н. эффекты урбанизации; позднее в литературе — от отраслевого разнообразия).

Если маршаллианские экстерналии в современной литературе рассматриваются, в сущности, так же, как о них писал А. Маршалл в 19 веке, то внимание к классификации Гувера — Улина обострилось благодаря работе Глейзера, Каллаль, Шейнкмана и Шлейфера «Growth in cities» [2]. В ней преимущества сосредоточения были разделены:

— на выгоды специализации при местной монополии (на языке авторов статьи это эффекты Маршалла — Эрроу — Ромера, т. е. авторов моделей экономического роста, где местное развитие подстегивает монополия на изобретения);

— экономию благодаря специализации, но при местной конкуренции (сюда Глезйер и пр. отнесли литературу, развитую вокруг идей о кластерах Портера);

— эффекты отраслевого разнообразия, или выгоды по Джейкобс.

Последние изначально связывались с тем, что полезные для производства какой-то отрасли идеи приходят прежде всего из других отраслей, а потому чем разнообразнее в каком-то месте отраслевой состав экономики, тем больше таких возможных источников, а значит и быстрее местный рост [3]. Но кроме указанного эффекта у отраслевого разнообразия есть и другие достоинства. Прежде всего т. н. портфельный эффект — страховка местной экономики от шоков [4].

2. Подходы к моделированию механизмов агломерационных эффектов

К оценке агломерационных эффектов, их разделению на виды главных подхода два.

Один восстанавливает регрессию микроэкономических функций, в уравнение которой среди независимых переменных помещают показатели специализации производства, отраслевого разнообразия и конкуренции фирм на указанной территории. Этот подход тесно связан с разделением выгод сосредоточения по образцу Гувера — Улина.

Второй подход восстанавливает регрессию для величины — показателя географического сосредоточения производств одной отрасли или пары отраслей. Независимые переменные при втором подходе — это показатели, которые по-разному отражают сходство отраслей, интенсивность и дешевизну связей между ними. В основе этого подхода лежит разделение внешних выгод сосредоточения по А. Маршаллу.

Современная литература, которая для измерения агломерационных эффектов и сравнения различных их механизмов, использует микроэкономические функции, восходит к статье Глейзера, Каллаль, Шейнкмана и Шлейфера «Growth in cities» [2].

Данными в их работе были показатели конкуренции и географической концентрации отраслей в 170 крупнейших городах США. Набор данных содержал информацию о занятости, зарплатах и количестве предприятий по отраслям для каждого округа в Соединенных Штатах. Мера местной конкуренции отрасли в городе определялась следующей формулой:

$$s_{ik} = \frac{\frac{e_{ik}}{\sum_i e_{ik}}}{\frac{\sum_k e_{ik}}{\sum_{ik} e_{ik}}}, \quad (1)$$

где s_{ik} — специализация города k на отрасли i,

e_{ik} — занятость в отрасли i в городе k.

Значение, превышающее единицу, означало, что в этой отрасли больше фирм по отношению к ее размеру в этом городе, чем в Соединенных Штатах. Это было бы свидетельством того, что отрасль в городе более конкурентная на местном уровне, чем на уровне США.

Глейзер и соавторы как меру разнообразия отраслей в городе, чтобы проверить теорию Джейкобс, использовали долю занятости в пяти крупнейших отраслях в городе. Чем ниже эта доля, тем разнообразнее город и тем быстрее должна расти рассматриваемая отрасль, согласно Джейкобс.

В итоге авторы составили следующую регрессию:

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + \beta_2 \alpha + \vec{\gamma} \cdot \vec{z} + \epsilon, \quad (2)$$

где $\ln(y)$ — логарифм отношения занятости в городских отраслях 1987 года к 1956 году,

$\ln(x)$ — логарифм отношения занятости в отраслях за пределами города 1987 года к 1956 году,

α — фиктивная переменная, обозначающая нахождение в южном регионе,

$\vec{z}$ — вектор нелогарифмированных объясняющих переменных, а именно: зарплаты в городской отрасли в 1956 году; занятости в городской отрасли в 1956 году; числа предприятий на одного работника в отрасли в городе по сравнению с тем же показателем по США в целом (обозначает уровень конкуренции); доли занятости в отрасли в городе по отношению к занятости по стране в целом; доли 5 крупнейших отраслей (для проверки эффектов Джейкобс);

$\beta_0, \beta_1, \beta_2, \vec{\gamma}$ — параметры;

ϵ — ошибка с нулевым ожиданием.

В рамках такого подхода сила различных механизмов агломерационных эффектов определяется сравнением оценок параметров в векторе $\vec{\gamma}$.

Альтернативный подход к сопоставлению разных видов агломерационных эффектов восходит к работам Эллисона и Глейзера 1997 г. [6], а также статье Эллисона, Глейзера и Керра 2009-2010 г. [7]. Далее для краткости будем говорить о подходе Эллисона — Глейзера — Керра. Идея подхода в том, чтобы сперва оценить сосредоточение пар отраслей (в какой степени они размещаются на тех же территориях или в каком-то ином смысле рядом), а затем сравнить полученные показатели с величинами, которые бы отражали иные виды сходства отраслей связанных с маршаллианскими экстерналиями:

1) интенсивностью обмена товарами между парой отраслей в производственной цепочке;
2) сходством требований к рабочей силе (чтобы можно было извлекать выгоду из большого местного рынка квалифицированного труда);
3) интенсивностью перетока знаний между отраслями (в таком случае местное сосредоточение указывало бы на обмен локализованными знаниями).

Две основных системы сосредоточения пары отраслей (т. н. коагломерации) задали работы Эллисона и Глейзера 1997 г. [6], а также Дюрантона и Овермана [8].

В статье 1997 года [6] Эллисон и Глейзер предложили меру для измерения степени сосредоточения отрасли, которая стала известна как индекс Эллисона — Глейзера:

$$\gamma_i \equiv \frac{\sum_j \left(\frac{\sum_i x_{ij}}{\sum_j x_{ij}} - \frac{\sum_i x_{ij}}{\sum_{i,j} x_{ij}}\right)^2 \Big/ (1 - \sum_m x_m^2) \; - H_i}{1 - H_i}, \qquad (3)$$

$$H_i \equiv \sum_{k=1}^{N_i} z_{ki}^2,$$

где γ_i — индекс Эллисона — Глейзера для отрасли i;

x_{ij} — число занятых в i-й отрасли территории j;

x_m — доля занятых территории m в числе занятых всех учтенных территорий;

H_i — индекс Херфиндаля для долей предприятий в i-й отрасли;

k — номер предприятия отрасли i,

N_i — число предприятий в отрасли i,

z_{ki} — доля занятых в фирме k отрасли i от общей занятости в отрасли i.

Показатель концентрации в формуле (3) рассчитывается только для одной отрасли. Однако Элллисон, Глейзер и Керр пошли дальше и стали рассматривать пары отраслей с помощью следующей меры:

$$\gamma_{ij}^c \equiv \frac{\sum_{m=1}^{M} (s_{mi} - x_m)(s_{mj} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}, \qquad (4)$$

где γ_{ij}^c — индекс Эллисона — Глейзера для пары отраслей i и j, мера их коагломерации;

M — число учтенных территорий;

s_{mi} — это доля занятых в i-й отрасли на территории m от общего числа занятых в i-й отрасли на всех территориях;

x_m — доля занятых территории m в числе занятых всех учтенных территорий [9,10] или среднее долей территории m в общем числе работников различных отраслей [7].

Таким образом Эллисон, Глейзер и Керр увеличили число наблюдений для модели регрессии, так как вместо одной отрасли сравнивается пара отраслей. Прежде Audretsch и Feldman, а также Rosenthal и Strange оценивали модель регрессии, где зависимой переменной была степень сосредоточения лишь отдельной отрасли [11,12]. Эллисон, Глейзер и

Керр, чтобы разделить источники выгод сосредоточения, использовали различия между парами отраслей: пара отраслей может быть похожей в одном отношении, но непохожей в другом. Предприятия одной пары отраслей могут располагаться близко друг к другу, так как у их отраслей общие требования к рабочей силе. Заведения другой пары отраслей могут предъявлять разные требования к работникам, но притягиваются друг к другу, чтобы экономить на перевозке товаров, так как одна отрасль пары — главный поставщик товаров для другой.

Что касается индексов Дюрантона — Овермана, то их отличие от индекса Эллисона — Глейзера состоит в том, что система Дюрантона — Овермана не использует дискретные пространственные единицы, ведь это делает расстояние между разными городами эквивалентными, даже если это и не является правдой. Вместо этого мера сосредоточения оценивается с помощью непрерывного индекса:

$$\widehat{K}_{ij}^{\,\mathrm{Emp}}(d) = \frac{1}{h\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} e(r)e(s)} \sum_{r=1}^{n_i}\sum_{s=1}^{n_j} e(r)e(s) f\left(\frac{d-d_{r,s}}{h}\right), \qquad (5)$$

где $\widehat{K}_{ij}^{\,\mathrm{Emp}}(d)$ — ядерная оценка плотности расстояния d между работниками заведений двух разных отраслей i и j;

h — коэффициент размытости гауссова ядра;

r — номер предприятия i-й отрасли;

s — номер предприятия j-й отрасли;

n_i и n_j — количество фирм в отрасли i и j соответственно;

$e(r)$ и $e(s)$ — число занятых на фирме r и s соответственно;

$d_{r,s}$ — расстояние между заводами r и s;

d — заданное расстояние;

f — гауссово ядро.

Если $e(r)$ и $e(s)$ принимаются равными единице, $\widehat{K}_{ij}^{\,\mathrm{Emp}}(d)$ превращается в $\widehat{K}_{ij}^{\,\mathrm{Ct}}(d)$ — ядерная оценка плотности расстояния d между заведениями (а не парой их работников) двух разных отраслей i и j.

Ядерная оценка плотности расстояний между парами работников заведений разных отраслей (или парами самих заведений) по формуле (5) предполагает, что известны координаты каждого предприятия. Эллисон, Глейзер и Керр такими данными не располагали, а потому рассчитали показатель по центроидам графств, в которых находились предприятия. Расстояние между любой парой предприятий в одном графстве принималось за 1 милю.

Чтобы ускорить вычисления, Эллисон, Глейзер и Керр в расчетах отрасли с более чем 1000 предприятий представляли случайной выборкой 1000 предприятий этой отрасли (без повторений). Критерием выбора коэффициента размытости гауссова ядра [h в формуле (5)] был минимум средней интегральной квадратичной ошибки [7].

Далее $\widehat{K}_{ij}{}^{\mathrm{Emp}}(d)$ или $\widehat{K}_{ij}{}^{\mathrm{Ct}}(d)$ рассчитываются снова и снова 1000 раз. Это дает подходящее для сравнений с реальным показателем эталонное распределение оценок. Чтобы его составить, на каждом из 1000 повторений из общего множества данных (предприятия всех отраслей) случайно отбираются предприятия так, чтобы общее число отобранных в две воображаемые отрасли предприятий совпадало с реальным числом предприятий в данных отраслях i и j (но не больше 1000). По данным пары (для двух воображаемых отраслей) выборок на каждом из 1000 повторений получается ядерная оценка плотности по формуле (5). Эталонное распределение ядерных оценок плотности расстояния было одно и то же для любой пары отраслей i и j.

Если для данного d оценка функции распределения была существенно больше 99-го квартиля соответственного показателя из эталонного распределения, Эллисон, Глейзер и Керр заключали, что пара отраслей глобально локализована. Если для данного d оценка функции распределения была существенно больше 99-го квартиля соответственного показателя из эталонного распределения, Эллисон, Глейзер и Керр заключали, что пара отраслей глобально рассеяна. Вместо интегрирования ядерной оценки плотности расстояний, Эллисон, Глейзер и Керр использовали простую сумму по расстояниям с шагом в 1 милю [формула (6)].

$$\begin{aligned} \Gamma_{ij}^{\mathrm{Emp}}(\bar{d}) &= \sum_{d=0}^{\bar{d}} \max\left(\widehat{K}_{ij}{}^{\mathrm{Emp}}(d) - K_{ij}{}^{\mathrm{UC}}(d), 0\right), \\ \Psi_{ij}^{\mathrm{Emp}}(\bar{d}) &= \sum_{d=0}^{\bar{d}} \max\left(K_{ij}{}^{\mathrm{LC}}(d) - \widehat{K}_{ij}{}^{\mathrm{Emp}}(d), 0\right), \end{aligned} \tag{6}$$

где $\Gamma_{ij}^{\mathrm{Emp}}(\bar{d})$ — показатель глобальной локализации работников отраслей i и j при пороговом расстоянии $\bar{d}$;

d — данное расстояние от 0 миль до $\bar{d}$ с шагом в 1 милю;

$\bar{d}$ — пороговое (максимальное для оценки локализации или рассеяния) расстояние;

$\widehat{K}_{ij}{}^{\mathrm{Emp}}(d)$ — ядерная оценка плотности расстояния d между работниками заведений двух разных отраслей i и j из формулы (5);

$K_{ij}{}^{\mathrm{UC}}(d)$ — 99-й процентиль эталонного распределения ядерных оценок плотности расстояния d между работниками заведений двух воображаемых отраслей i и j;

$\Psi_{ij}^{\mathrm{Emp}}(\bar{d})$ — показатель глобального рассеяния работников отраслей i и j при пороговом расстоянии $\bar{d}$ ($\Psi_{ij}^{\mathrm{Emp}}(\bar{d})$ 0, если $\Gamma_{ij}^{\mathrm{Emp}}(\bar{d})$ не 0);

$K_{ij}{}^{\mathrm{LC}}(d)$ — 1-й процентиль эталонного распределения ядерных оценок плотности расстояния d между работниками заведений двух воображаемых отраслей i и j.

Чтобы оценить, в какой степени предприятия и отрасли торгуют друг с другом, Эллисон, Глейзер и Керр ввели следующую переменную:

$$InputOutput_{ij} = \max\left(\max\left(\frac{x_{ij}}{\sum_k x_{ik}}, \frac{x_{ji}}{\sum_k x_{jk}}\right), \max\left(\frac{y_{ij}}{\sum_k y_{ik}}, \frac{y_{ji}}{\sum_k y_{jk}}\right)\right), \quad (7)$$

где $InputOutput_{ij}$ — показатель интенсивности обмена товарами между отраслями i и j;

i, j, k — номера отраслей;

x_{ij} — объем закупок отраслью i товаров отрасли j;

y_{ij} — объем продаж отрасли i отрасли j.

Эти доли рассчитываются относительно всех поставщиков и клиентов.

Для оценки важности сходства требований пары отраслей к рабочей силе (а значит, и возможностей, которые они получают при размещении на той же территории с крупным рынком специалистов нужной квалификации) Эллисон, Глейзер и Керр использовали коэффициент корреляции долей работников разных профессий в двух отраслях.

Для оценки интенсивности обмена знаниями между отраслями, а значит, возможных выгод от локальных перетоков знаний, Эллисон, Глейзер и Керр ввели следующую меру:

$$Tech_{ij} \equiv \max\left(\max\left(\frac{x_{ij}}{\sum_k x_{kj}}, \frac{x_{ji}}{\sum_k x_{ki}}\right), \max\left(\frac{x_{ij}}{\sum_k x_{ik}}, \frac{x_{ji}}{\sum_k x_{jk}}\right), \right), \quad (8)$$

где $Tech_{ij}$ — показатель интенсивности обмена знаниями между отраслями i и j;

i, j, k — номера отраслей;

x_{ij} — оценка суммы расходов на НИОКР, понесенных отраслью i, которые принесли выгоду отрасли j (на основе опроса специалистов о полезности для разных отраслей патентов из крупной выборки) (элемент матрицы в таблице 20.1 в тексте Шерера 1984 г. [13]).

Итоговое уравнение регрессии в подходе Эллисона — Глейзера — Керра (основная спецификация) принимает такой вид:

$$Coagg_{ij} = \alpha + \beta_{\mathrm{NA}} Coagg_{ij}^{\mathrm{NA}} + \beta_{\mathrm{L}} LaborCorrelation_{ij} + \beta_{\mathrm{IO}} InputOutput_{ij} + \beta_{\mathrm{T}} Tech_{ij} + \varepsilon_{ij} \quad (9)$$

где $Coagg_{ij}$ — это мера коагломерации предприятий пары отраслей i и j: или индекс Эллисона — Глейзера, или «индекс Дюрантона — Овермана» [1];

$LaborCorrelation_{ij}$ — коэффициент корреляции долей работников разных профессий в двух отраслях (отражает выгодность доступа к постоянному рынку похожего квалифицированного труда);

$InputOutput_{ij}$ — показатель интенсивности обмена товарами между отраслями i и j из формулы (7);

$Tech_{ij}$ — показатель интенсивности обмена знаниями между отраслями i и j из формулы (8);

$Coagg_{ij}^{\mathrm{NA}}$ — основанный на оценках по формуле показатель (аналогичный $Coagg_{ij}$) гипотетической коагломерации отраслей, если бы их размещение определяли только «естественные преимущества»;

ε_{ij} — ошибка регрессии с нулевым ожиданием.

Оценки параметров этой регрессии для США подтвердили важность всех маршаллианских экстерналий, но самым важным (по величине оценки соответствующего параметра) фактором была объявлена экономия на издержках за счёт близости к поставщикам ресурсов или конечному потребителю, то есть переменная $InputOutput_{ij}$. Вместе $LaborCorrelation$, $InputOutput$ и $Tech$ объяснили бóльшую часть дисперсии значений показателей коагломерации, чем $Coagg_{ij}^{\mathrm{NA}}$: агломерационные эффекты сильнее влияли на размещение отраслей, чем «естественные преимущества».

[1] Эллисон, Глейзер и Керр не объясняют, что именно они понимают под таким индексом. Показателем могли служить: $\widehat{K}_{ij}^{\mathrm{Emp}}(d)$ из формулы (6); значение соответствующей ей функции распределения для того же порогового расстояния d; глобальный индекс локализации

3. Разработка подхода к идентификации фактических границ агломераций в России с учетом пространственного распределения экономической активности

Для успешного применения моделей, которые изучают выгоды сосредоточения с помощью микроэкономических функций, нужно, чтобы территориальные единицы, по которым делается расчет степени местной специализации или отраслевого разнообразия, были в существенной мере самостоятельными, как это предполагает сама теория, где рост отраслей города увязывается со свойствами самого этого города, но не свойствами всех других городов.

Именно поэтому в рамках работы предприняты попытки разработки подходящего для России способа выделения агломераций и/или кластеров построек/производителей.

3.1. Формирование базы данных для проведения делимитации границ агломераций в России

Координаты расположения зданий и их тип взяты из базы данных Open Street Maps [14] по всем 85 регионам. Срез произведен на момент 13 февраля 2022 г. по группам объектов с тегом «building=*», из которых были убраны объекты, не представляющие интереса и вносящие шум: дубликаты, нерелевантные объекты (например, диспетчерские вышки, ветроуказатели, антенны, столбы, трубопроводы, ограды, турникеты, шлагбаумы и т. д. Около 30 % зданий имеют указанное назначение (завод, магазин, школа, жилой дом и т. д.). Собранные данные позволяют применить алгоритм кластеризации DBSCAN, о котором будет рассказано далее.

3.2. Формирование подхода к проведению анализа, разработка архитектуры модели

Алгоритмы кластеризации широко применяются для решения задач по идентификации. Однако применение алгоритмов кластеризации к большим пространственным базам данных предъявляет следующие требования:

— знания в предметной области для корректного определения входных параметров,
— возможность обнаружения кластеров произвольной формы,
— высокая эффективность на больших объемах данных (от нескольких тысяч объектов и более).

Большая часть алгоритмов кластеризации не удовлетворяет сразу всем трем требованиям. На этом фоне выгодно выделяется алгоритм кластеризации DBSCAN, первоначально описанный Ester и др. в работе 1996 г. [15]. DBSCAN, или плотностный алгоритм пространственной кластеризации с шумами, – неконтролируемый алгоритм машинного обучения, который применяется для классификации неразмеченных данных.

Поведение модели определяется несколькими параметрами [16]:

— Eps: две точки считаются соседями, если расстояние между ними меньше порогового значения Eps;
— $MinPts$: минимальное количество соседей, которое должна иметь данная точка, чтобы образовать кластер. Важно отметить, что сама точка при этом входит в состав $MinPts$;
— метрикой, используемой при расчете расстояния между точками в массиве (например, наиболее часто используемое евклидово расстояние).

Алгоритм начинает работу с вычисления расстояния между каждой точкой и всеми прочими точками. В итоге точка может быть отнесена к одной из трех категорий):

1) точка-ядро — точка, вокруг которой есть $MinPts$ точек, расстояние до которых относительно точки ниже порогового значения, определенного Eps;
2) пограничная точка — точка, которая не находится в непосредственной близости ото всех точек в области $MinPts$, но близка к одной или нескольким точкам-ядрам. Пограничные точки включаются в кластер ближайшей точки-ядра;
3) точка-шум — точка, расстояние от которой до точек-ядер больше Eps. Точки-шумы игнорируются и не являются частью какого-либо кластера.

От того, как будут выбраны значения Eps и $MinPts$, зависит, насколько корректным получится итоговый расчет. Rahmah и Sitanggang [17] в 2016 г. предложили подход к автоматическому определению оптимального значения Eps, которое располагается в точке максимальной кривизны.

Суть DBSCAN состоит в определении того, достаточно ли близко расположенное минимальное количество точек друг к другу, чтобы считаться частью одного кластера. При этом алгоритм чувствителен к масштабу, т. к. Eps фиксирован для всей выборки данных.

Как отмечается у Gao [18], DBSCAN не очень хорошо работает с плотностями с высокой дисперсией. В случае использования географических данных это может создавать дополнительные трудности, особенно когда речь идет о многомиллионных выборках. Особенности работы алгоритма на данных с разной плотностью хорошо иллюстрирует пример кластеризации зданий в Казани и ее окрестностях. При некоторых комбинациях параметров в кластер попадет не сама Казань, а облако близко расположенных точек – большой массив ИЖС на севере, на который приходится около 21% всех точек города.

4. Эмпирическое тестирование разработанного подхода к делимитации границ агломераций в России с помощью методов машинного обучения, анализ и интерпретация полученных результатов

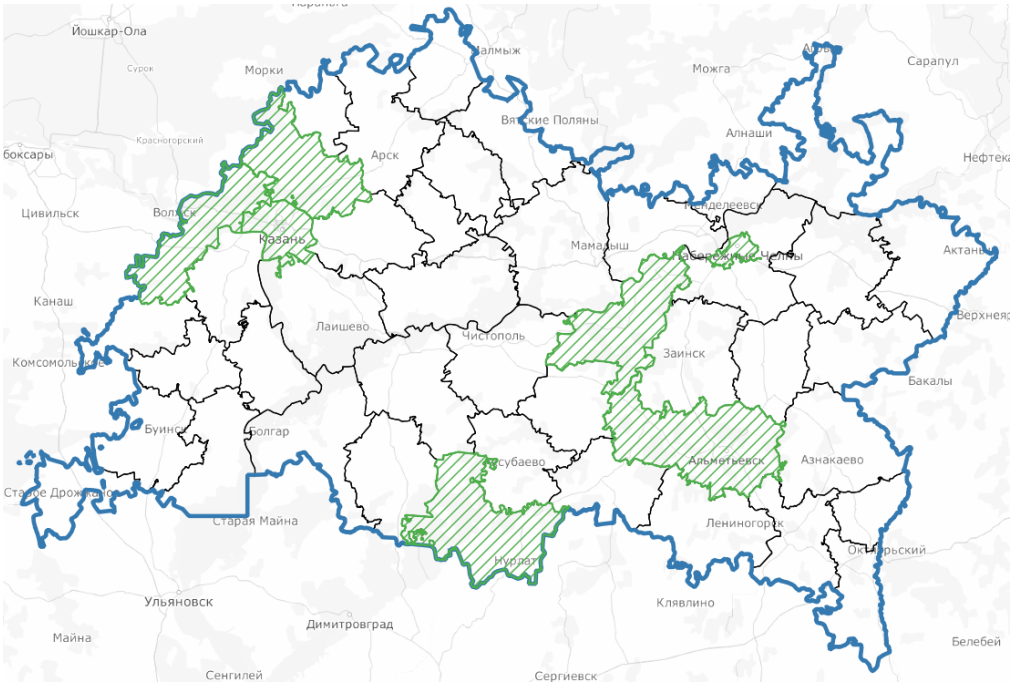
4.1. Кластеризация без учета этажности с использованием пула параметров

Далее в работе будут рассматриваться кластеры, полученные путем реализации алгоритма DBSCAN со всеми комбинациями $eps = [5, 10, 15]$ км и $MinPts = [500, 1\,000, 1\,500, 2\,000, 2\,500, 3\,000]$ зданий, то есть будет рассмотрено 18 пар входящих параметров, 18 результатов кластеризации для каждого региона.

Все здания, попавшие в тот или иной кластер, соотнесены с муниципальным образованием, в котором они расположены (уровень детализации – муниципальные районы и городские округа). Чтобы выяснить, следует ли считать муниципальное образование частью кластера, необходимо высчитать, какая доля точек в нем была отнесена к этому кластеру. Для этого нужно найти сумму точек, отнесенных к каждому кластеру внутри каждого муниципального образования, в котором есть точка кластера. Пороговое значение в 50% означает, что муниципалитет привязывается не более чем к одной агломерации. Получившиеся таким образом кластеры могут состоять из любого числа муниципальных образований (даже из одного) и иметь любой состав (например, только муниципальные районы без городских округов).

Для выявления наиболее устойчивых кластеров использовался следующий подход: если муниципалитет попадал в кластер в своем регионе в 75% и более случаях (т. е. в как минимум в 13 из 18 кластеров), то его можно отнести к устойчивым. Соответственно, выборка таких муниципальных образований и формирует устойчивый кластер (устойчивые кластеры) региона. По запросу авторами может быть представлен перечень муниципалитетов с указанием принадлежности к региону и частотой попадания в кластеры.

Примечательно, что при подобном подходе зачастую устойчивы даже небольшие по численности населения кластеры. Это хорошо видно на примере упоминавшейся ранее Республики Татарстан (рисунок 1, муниципалитеты, формирующие кластеры, отмечены зеленым): малонаселенные (по сравнению с Казанью и Набережными Челнами) города Нурлат, Зеленодольск и Нижнекамск устойчивы при широком выборе пар параметров.



Примечание — Источник: составлено авторами.

Рисунок 1. Устойчивые кластеры на примере Республики Татарстан

4.2. Сравнение результатов с потенциальными границами агломераций из предыдущего исследования

В НИР 2020 г. «Разработка научно-методологических подходов к идентификации фактических границ агломераций в России с учетом пространственного распределения экономической активности» [19] был использован комбинированный подход: выделение ядер агломераций по методологии OECD. На первом этапе производится идентификация ядер на сетке данных о плотности населения. В ядро входят ячейки с плотностью населения выше 1 000 чел/км^2, при этом общая численность получаемого пятна более 100 тыс. чел. При переходе от сетки к административно-территориальному делению используется правило: в ядро входят муниципалитеты, более 50% населения которых размещено в отобранных областях [19]. Дальнейшее следование методологии ни тогда, ни сейчас не представляется возможным ввиду отсутствия данных о межмуниципальной ежедневной (маятниковой) миграции, поэтому рассчитаны полуторачасовые изохроны.

Чтобы сравнить результаты кластеризации DBSCAN и модифицированной методики OECD (OECDм), в каждом регионе отобраны муниципалитеты, которые попадают в области OECDм, то же самое сделано для каждого из пороговых значений DBSCAN. Затем определено, при каком из пороговых значений DBSCAN в область попадает максимально близкое число муниципалитетов, что и в случае OECDм.

Если сравнивать методики сразу по двум параметрам, то результат такого сравнения получается разнообразен: в разных регионах с метрикой OECDм схожи разные пороговые

значения DBSCAN, но в основном это $eps = 15, MinPts = 500$ (20 регионов) и $eps = 15, MinPts = 1000$ (9 регионов). Поскольку сравнивалось только число муниципалитетов, то столь частое совпадение с парой $eps = 15, MinPts = 500$ формально выглядит логично, потому что полуторачасовая изохрона охватывает значительную область в методике OECDм, равно как и такая пара входящих параметров позволяет учесть даже небольшие города-кластеры (то есть в результате получается обширная область, самая большая из всех пар параметров). Однако качественное различие между двумя методиками существенно: OECDм располагает полученную область вокруг крупного города (чаще всего – центра региона), а DBSCAN – маленькие кластеры по всей территории. Если рассматривать параметры по отдельности, то для eps видно, что в самое распространенное значение – 5 или 15 км, $eps = 10$ км характерно только для 8 регионов. Что касается числа зданий $MinPts$, то для большинства регионов достаточно порога в 500 и 1 000 зданий для того, чтобы быть похожим на метрику OECDм. Для всех примеров соответствия параметров сложно выделить пространственную закономерность, но можно предположить, что за счет лучшей транспортной связанности в европейской части страны на методику OECDм в большей степени похожи обширные одиночные пятна кластеров, образующиеся при $eps = 15$ км, то есть изохрона позволяет их «догнать», в то время как в восточной части страны ее протяженности для этого не хватает.

5. Разделение отбора предприятий и выгод сосредоточения производства на примере Калининградской области

Неравномерность размещения производства — главное наблюдения экономической географии и отправная точка для ее изысканий. Обычно сосредоточение производства, возникновение скоплений производителей (городов, агломераций, территориально-производственных комплексов, кластеров и пр.) объясняют тем, что сосредоточение, если взвесить все за и против, в итоге выгодно.

Хотя модели, которые объяснили возникновение и развитие городов (шире — скоплений хозяйственной деятельности) внешними выгодами (англ. positive externalities, external economies) сосредоточения, стали весьма успешными (пример — т. н. новая экономическая география), само измерение таких выгод, проверка выводов моделей наталкивается на статистические трудности. Прежде всего речь о том, что, если сосредоточение выгодно, то мы должны наблюдать более высокую норму прибыли, производительность, зарплату, ренту и т. п. в более крупных, плотных скоплениях производителей.

Тем не менее, выгоды сосредоточения — это не единственная причина, которая может объяснить такие наблюдения. Так, у производителей внутри кластера показатели могут быть лучше из-за более жесткого отбора. На то, что в кластерах выше смертность предприятий, указывали в тематическом литературном обзоре Маккан и Фолта [4]. Если пренебречь отбором среди производителей лучших благодаря конкуренции в кластерах, оценки величины выгод сосредоточения будут содержать систематическую ошибку.

Чтобы решить эту проблему, мы используем подход, предложенный Комбом и др. [20]. Суть этого подхода в том, что отбор и выгоды сосредоточения по-разному изменяют распределение производительности предприятий. Отбор делает его усеченным слева (в подходе Комба и др. из микроэкономических начал модели следует, что речь о распределении не самой производительности производителей, но ее логарифма), отсекает слева долю предприятий *S*, производительность которых оказалась слишком низкой, чтобы оставаться на рынке:

$S_i \equiv 1 - G(\bar{h}_i)$, где S_i — доля предприятий территории i, которые не выдержали конкуренции; $G(\bar{h}_i)$ — функция распределения предельно возможных (чтобы был ненулевой спрос на товар) предельных затрат.

Выгоды сосредоточения в подходе Комба и др. искажают распределение производительности предприятий иначе: или сдвигают его вправо на величину *A*, или растягивают правый хвост распределения, что отражает параметр *D*:

$A_i \equiv \ln a\left(N_i + \delta \sum_{i \neq j} N_j\right)$, $0 \leq \delta \leq 1, a(0) = 1$, $a' > 0$, $a'' < 0$, где i, j — номера территорий, а — функция ценности единицы труда, если лучшие предприятия не получают бóльших преимуществ от сосредоточения; N — число работников/потребителей; δ определяет, как сильны связи между территориями по сравнению со связями внутри территории;

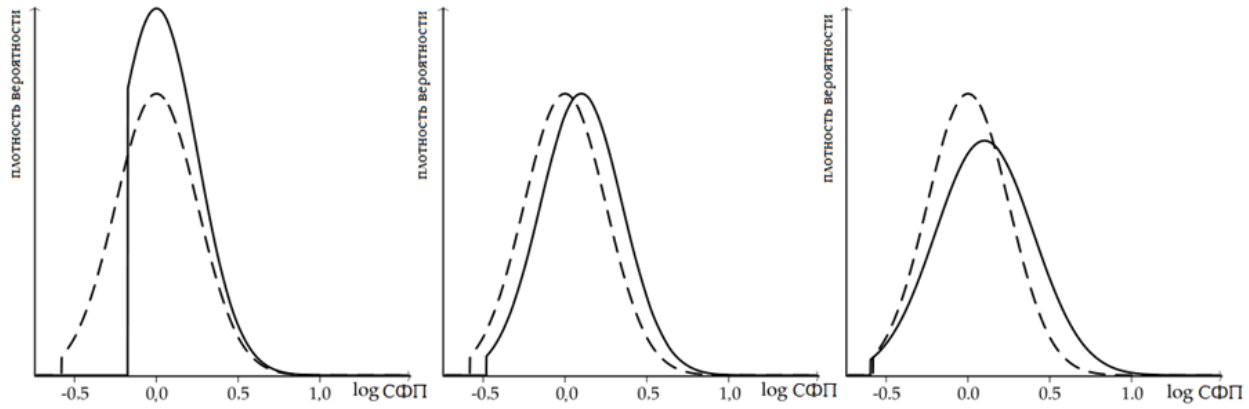
$D_i \equiv \ln d\left(N_i + \delta \sum_{i \neq j} N_j\right), 0 \leq \delta \leq 1, d(0) = 1$, $d' > 0$, $d'' < 0$, где D_i — величина, делающая ценность единицы труда больше, если она используется более успешным предприятием, так что ценность единицы труда определяется так: $\frac{\ln a\left(N_i + \delta \sum_{i \neq j} N_j\right)}{h^{D-1}}$, где h — предельные издержки предприятия.

Если смысл сдвига всего распределения интуитивен: все производители внутри их скопления получают выгоды от сосредоточения, то смысл растяжения параметром *D* несколько сложнее: более производительные предприятия получают бóльшие преимущества от сосредоточения, чем менее производительные. Источники таких неравных преимущества могут быть разными: из-за большей восприимчивости к передовым технологиям благодаря найму более квалифицированных работников (рынок которых в крупном городе больше), лучшей организации труда и пр.

Пусть $A \equiv A_i - DA_j$, $S \equiv \frac{S_i - S_j}{1 - S_j}$, $D \equiv \frac{D_i}{D_j}$. Если $\mathrm{A} > 0$, агломерационные эффекты улучшают прибыльность всех предприятий. Если $\mathrm{A} > 0$, $\mathrm{D} > 1$, более прибыльные предприятия извлекают бóльшую выгоду от сосредоточения. Если $\mathrm{S} > 0$, то на плотных территориях сильнее отбор предприятий.

Задача метода Комба и др. (2012) — восстановить с помощью только этих 3 параметров (*A, D, S*) из наблюдаемого распределения производительности предприятий одной территории распределение для другой, чтобы минимизировать квадрат разности квантилей между наблюдаемым и восстановленным распределением логарифма производительности территории.

Графически смысл перехода от одного распределения к другому показывает рисунок 2.



Примечание — Источник: составлено автором на основе рисунков в [20].

Рисунок 2. Изменение исходного распределения с помощью параметров *A, D, S*. Слева направо: $A = 0,\ D = 1,\ S > 0;\ A > 0,\ D = 1,\ S > 0;\ A > 0,\ D > 1,\ S > 0$ (следствия отбора, только сдвига всего распределения, а также и сдвига, и растяжения распределения соответственно)

Данными для применения описанного метода стали сведения обо всех коммерческих организациях в 2018 г. в Калининградской области без территориально обособленных подразделений и с известными / ненулевыми издержками и выручкой. Выбор 2018 г. объясняется тем, что это последний год, для которого данные бухгалтерской отчетности организаций размещались в открытом доступе (с 2019 г. — платно от ФНС). Калининградская область была выбраны для пробы метода, так как не имеет границы с другими регионами России, что позволяет исключить неучтенное влияние соседних российских городов и пр. Кроме того, для небольшой территории существенно легче собрать сведения о наличии и размещении территориально обособленных подразделений: у нас не было возможности получить разово полную базу для всех организаций, а потому мы полагались на автоматизацию запросов (не чаще одного запроса примерно в 7 с для одного IP-адреса) к порталу Росстата https://websbor.gks.ru/. Исключение подразделений потребовалось, чтобы можно было однозначно привязать наблюдаемые показатели предприятия к одному из двух распределений в формулах выше (i или j). Источником сведений об адресах предприятий и их организационно-правовой форме (для отбора только коммеерческих организаций, стремящихся получить прибыль, как это и подразумевает модель Комба и др.) была выгрузка ЕГРЮЛ по состоянию на 1 января 2020 г., откуда были отобраны организации, действовавшие в 2018 г. Мерой производительности для наших оценок служила прибыльность. Применение этого показателя освобождает от существенных эконометрических и содержательных трудностей, возникающих при попытке оценить совокупную факторную производительность по

стоимостным величинам предприятий в смешанных ценах. Прибыльностью было отношение выручки к расходам по обычной деятельности.

Мы рассмотрели два способа выделения распределений по географическому признаку:

1) предприятия округов с плотностью жителей выше медианы по области против предприятий прочих округов (16 007 предприятий);
2) предприятия (не в административно-территориальных районах, но только в городах — в силу сложности поиска координат) в выделенном с помощью алгоритма DBSCAN [1] кластере против предприятий вне кластера (10 558 предприятий).

Первый подход — традиционный, он опирается на административно-территориальное деление. Этот же метод использовали Комб и др. в оригинальном исследовании. Второй же способ вытекает из сомнения в том, что муниципальные образования подходят, чтобы судить о выгодах сосредоточения производства. Модели в географической экономике подразумевают самостоятельные территориальные единицы, часто именуемые кластерами или агломерациями (скоплениями производителей). Самостоятельность таких единиц означает, что они не влияют друг на друга прямо, а потому территории со смежными границами, такие как муниципальные образования, плохо подходят для проверки теорий. Кроме того, модели обычно исходят из того, что производители максимизируют прибыль, при этом единицей, которая принимает нужные решения, в моделях выступает фирма.

Для работы алгоритма DBSCAN исследователь задает два параметра: наименьшее число точек в кластере (принималось равным 3 000) и максимальное расстояние между точками одного кластера (принималось равным 5 км). Преимущество DBSCAN перед другими методами выделения агломераций/кластеров в том, что этот метод умеет связывать кластеры через плотно застроенные «перемычки».

Для получения нужных для DBSCAN координат предприятий использовался локальный сервер Nominatim на пространственных данных OpenStreetMap. Найденные координаты были избирательно выверены вручную по ~100 предприятиям. Ошибок не было. С помощью OSM и Nominatim геокодировано 81,2 % предприятий *в городах области.*

Результаты выделения кластера калининградских предприятий показывает рис. 3. На этих рисунках видно, что предприятия размещены компактнее, чем жилая застройка. Размер кластера предприятий меньше, чем кластера жилищ: меньше выходы из Калининграда в Гурьевский городской округ.

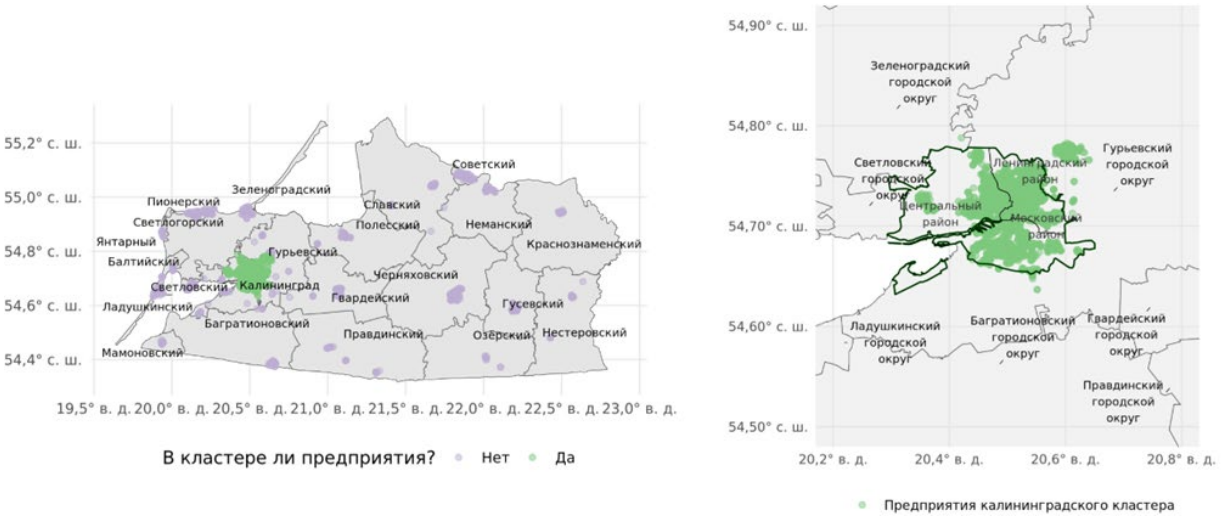


Рисунок 3. Сравнения границ кластера предприятий в Калининградской области с административными границами

Результаты оценок приводит таблица 1. В скобках полученная с помощью бутстрепа (100 выборок с повторением) оценка квадратичной ошибки. Звездочка — если оценка параметра лежит вне оцененного 95%-ного интервала. Для A и S статистически значима оценка, отличная от 0, для D — отличная от 1. Псевдо-R^2 показывает, какую часть среднего квадрата разности квантилей между распределением логарифма прибыльности предприятий в плотных и неплотных округах объясняют три параметра. Указанное в таблице число предприятий меньше, чем общее число доступных наблюдений, так как отсекался 1 % снизу и сверху распределения, чтобы убрать выбросы.

Таблица 1

Результаты оценок

Что сравнивалось	A	D	S	Число наблюдений	Псевдо-R^2
Округá с разной плотностью жителей	0,071* (0,013)	1,077 (0,092)	0,004 (0,006)	15 689	0,570
Предприятия в калининградском кластере или вне его	−0,011 (0,010)	0,868* (0,063)	−0,011* (0,004)	10 347	0,920

Главные выводы таковы:

— на уровне округов сосредоточение выгодно всем (примерно на 7 %), нет признаков отбора, преимуществ для лучших;

— но в калининградской агломерации (калининградском кластере) в сравнении с другими территориями отбор (конкуренция) предприятий меньше, что указывает на неблагоприятный отбор: предприятия с большими удельными издержками извлекают бóльшую выгоду от расположения в кластере.

Второй вывод можно объяснить так. Неэффективные предприятия из-за ограниченности ресурсов вынуждены либо использовать худшую технологию, либо копировать технологии более передовых предприятий. Возникают экстерналии, которые, однако, полезны непроизводительным, а не успешным предприятиям. Успешные же предприятия из-за таких экстерналий теряют стимулы сосредоточиваться в определённой местности. Этот вывод согласуется с выводами о том, что у фирм низкая выживаемость в кластерах и что крупные и успешные фирмы могут быть не заинтересованы в агломерации.

6. Применение метода Эллисона — Глейзера — Керра к полному кругу организаций в России

Основой для сравнения силы различных агломерационных эффектов в России стал метод Эллисона — Глейзера — Керра. Отличительной особенностью, однако, стал отказ от применения индексов Эллисона — Глейзера или Дюрантона — Овермана для измерения коагломерации пар отраслей. Вместо этого мы использовали более новые показатели, предложенные Ховард, Ньюман и Тарпом. Их индексы изначально созданы для работы с дискретными территориальными единицами, а не координатами, удобны при счете числа производителей, а не только рабочих, сравнивают наблюдаемое распределение с ожидаемым, чтобы оценить необычность наблюдений, имеют жесткую шкалу и просты в интерпретации.

Основа подхода — оценка наблюдаемого сосредоточения пары отраслей. Вслед за Ховард, Ньюман и Тарпом будем называть этот показатель «соразмещения» CL-индексом (CL — от англ. co-locaction). Этот индекс отвечает на вопрос: в какой степени две отрасли размещаются в тех же местах? В сущности, речь о частоте (вероятности), с которой предприятия пары отраслей находятся в том же месте:

$$CL_{ij} = \frac{\sum_{k=1}^{p} \sum_{l=1}^{q} C_{kl}}{p \times q}, \quad (10)$$

где CL_{ij} — это CL-индекс для пары отраслей i и j;

p — число организаций (желательно — вообще всех производителей) в отрасли i;

q — число организаций в отрасли j;

C_{kl} — величина, равная 1, если организации k и l (соответственно из отрасли i и отрасли j) располагаются на той же территории (в том же муниципальном районе, городском округе или на той же внутригородской территории города федерального значения), или 0, если организации k и l располагаются на разных территориях.

Из формулы (10) ясно, что CL-индекс определен на отрезке от 0 (производство пары отраслей полностью рассредоточено по разным территориям) до 1 (все производства пары отраслей сосредоточены на одной территории). Формула также указывает и на явные слабые стороны индекса:

как и индекс Эллисона — Глейзера, CL-индекс Ховар — Ньюман — Тарпа никак не может учесть близкого размещения производств пары отраслей, если они находятся на пусть и смежных или близких, но формально (по коду ОКТМО) разных территориях;

индекс, в отличие от показателей системы Дюрантона — Овермана, чувствителен к проведению границ территорий [англ. modifiable unit area problem (MAUP)]. Так, очевидно, что чем более дробное территориальное деление мы используем, тем меньше вероятность получить CL-индекс, близкий к 1.

Последний изъян, однако смягчает второй показатель системы Ховард — Ньюман — Тарпа, так как этот показатель сравнивает наблюдаемый CL-индекс с ожидаемым, исходя из общего размещения производств пары отраслей. Речь о так называемым XCL-индексе (XCL — от англ. excess co-location, т. е. это мера «избыточности соразмещения» пары отраслей). Чтобы получить эту меру отклонения степени размещения отраслей в одном месте от общего сосредоточения двух отраслей в целом, используется бутстреп. Для каждой пары отраслей i и j есть $p + q$ организаций. Суть в том, чтобы из этого множества организаций случайно отобрать [2] сперва p организаций и отнести их к отрасли i, затем из того же множества $p + q$ организаций случайно отобрать q организаций и приписать их отрасли j [3] и рассчитать CL-индекс для этих воображаемых способов размещения отраслей i и j. Такую операцию следует проделать как можно большее число раз [4], чтобы точнее оценить ожидание CL-индекса — среднее от оцененных по бутстреп-выборкам CL-индексов. Искомый XCL-индекс получается вычитанием из наблюдаемого CL-индекса оценки его ожидания. Показатель будет между -1 и +1. Положительная величина XCL-индекса означает, что организации пары отраслей располагаются в тех же местах чаще ожидаемого (исходя из общего размещения производства).

Для переложения на российские условия метода Эллисона — Глейзера — Овермана использовали такие данные:

1. Срез ЕГРЮЛ на 01.01.2020:

— отобраны организации, действовавшие на 01.01.2020, удалены повторения из-за ошибок регистрации и пр. (тот же ИНН). Всего более 3,7 млн организаций**;**

— все коды ОКВЭД приведены к ОКВЭД-2 (перекодирование из, ОК 029-2001 (КДЕС Ред. 1 и 1.1): охват всех отраслей, любая дробность (~79 % с 3 цифрами, 66 % с 4, но с 2012 г.);

— кода ОКВЭД обычно нет для ликвидированных организаций → только на 01.01.2020;

[2] Мы использовали простую случайную выборку, то есть отбирали элементы множества с повторением. Создатели индексов умалчивают о том, каким способом брали выборку.

[3] p и q при этом — это наблюдаемое число организаций в отрасли i и j соответственно.

[4] Мы, как и Ховард, Ньюман и Тарп, использовали 50 итераций, то есть работали с сотней (50×2) выборок.

— учтены территориально обособленные подразделения (филиалы, представительства)

2. Статистический регистр хозяйствующих субъектов Росстата на 01.01.2020, благодаря которому удалось собрать коды ОКТМО. Все они были приведены к состоянию на 01.01.2020. Был взят уровень района. Всего 2 610 территориальных единиц.
3. Коды статистики Росстата (http://websbor.gks.ru/). Этот ресурс позволил учесть территориально обособленные подразделения и найти их ОКТМО.

Первое, что бросается в глаза в полученных оценках коагломерации — это устойчивость к выбору отраслевого деления вывода о том, что в России пары отраслей обычно *рассредоточены* относительно друг друга. Вывод о рассредоточенности производств различных отраслей противоречит привычным в литературе заключениям о сосредоточении отраслей, необычной близости их размещения. Такое расхождение мы склонны связывать по меньшей мере для российских оценок с тем, что наш анализ — наиболее полный по охвату организаций. Впервые мы представили анализ для российских производителей, который на указанную дату (1 января 2020 г.) охватил все организации, у которых в ЕГРЮЛ был указан код вида деятельности.

Так как число работников разных профессий по отраслям неизвестно (данные о средних и крупных организациях по классификатору занятий слишком общие), использовались для оценки сходства требований к рабочей силе данные вакансий из «Работы России» (trudvsem.ru). Эта база предоставила для анализа 1,5 млн вакансий (последние версии) для организаций, действовавших на 01.01.2020. При оценке корреляции долей вакансий для разных профессий мы делали поправку на число рабочих мест. Полезно, что код профессии (классификатор ОКПДТР давно не искажали изменения) был указан для 84,2 % вакансий. Всего в базе были вакансии по 5 тысячам кодов. При этом в среднем приходилось 300 вакансий на каждую вакансию. Для организаций был указан ОГРН, что дало привязку к основной базе данных на основе ЕГРЮЛ, статистического регистра хозяйствующих субъектов, кодов статистики Росстата.

Для оценки интенсивности связей между поставщиками и потребителями использовалась гармонизированная матрица затрат-выпуска для России, представленная ООН по состоянию на 2018 г. Для расчета сходства отраслей в технологической сфере использовалась матрица Шерера. Из-за согласования устаревших отраслевых кодов SIC, на которых основана матрица Шерера с ОКВЭД-2, произошла вынужденная агрегация отраслей. В итоге для работы осталось 19 отраслевых групп.

Для выяснения силы связи с сосредоточением пар отраслей интенсивности обмена между ними по матрице затрат-выпуска и по матрице Шерера, а также сходства требований пар отраслей к рабочей силе, мы оценивали параметры простых линейных уравнений без

каких-либо преобразований зависимых переменных. Подобный подход возможен, так как у переменных удобные и понятные шкалы значений. У всех величин значения бывают от 0 до 1. Исключение составляют показатель сходства требований к рабочей силе (коэффициент корреляции принимает значения от -1 до +1, хотя в нашем ряде оценок почти все величины положительные) и XCL-индекс (также от -1 до +1). Однако проще просто учитывать эти особенности при интерпретации результатов, чем полагаться на зависящие от конкретного набора данных преобразования, например к z-величинам. Существование отрицательных величин — одна из причин отказа от логарифмического преобразования. Применение же более сложных преобразований [главный кандидат — преобразование Ё — Джонсона (англ. Yeo — Johnson power transformation), которое может сделать распределение величин более симметричным и близким к нормальному, но, в отличие от преобразования Бокса — Кокса, не требует, чтобы исходные величины были положительными [21]], но их применение бы еще более усложнило интерпретацию оценок параметров при трех факторах. Кроме того, при подборе оптимального параметра самого преобразования расходовались бы степени свободы, но такую их растрату сложно было бы учесть [22].

Чтобы сделать сравнение важности трех маршаллианских экстерналий еще более явным, мы использовали полный перебор всех возможных спецификаций модели линейной регрессии, возможных при упрощении уравнения с тремя исходными переменными и всеми их парными произведениями. Такие произведения нужны, чтобы учесть, как влияние на сосредоточение производств одного механизма агломерационных выгод зависит от другого. При таком подходе всего возможно 18 уравнений, включая уравнение с одной только константой (англ. intercept).

Отбор лучших моделей (далее мы представим 5 лучших) мы производили по критерию Акаике.

Необычно высокие оценки сосредоточения получают пары отраслей с малым числом организаций. Такие ненадежные из-за числа участвовавших в расчете показателей пары отраслей могли бы исказить оценки параметров уравнений регрессии. Чтобы учесть эту возможность и проверить устойчивость оценок, мы использовали, кроме обычного эстиматора МНК, также эстиматор, который исходил из минимизации взвешенной суммы квадратов остатков. Весами при этом было число организаций в меньшей по числу организаций отрасли из данной пары отраслей. Взвешивание квадрата остатка происходило простым умножением на соответствующую величину веса пары. При таком подходе подбор параметров должен был проходить так, чтобы как можно лучше описывать более надежные измерения, в которых участвовало большое число организаций.

Указанный способ оценки, однако, подходит, лишь чтобы сравнить важность различных маршаллианских экстерналий в среднем для разных отраслей. Чтобы выяснить, какие силы особенно важны для отдельного вида деятельности, описанный выше метод не подходит. Всё дело в том, что в его уравнения трудно добавить фиктивные переменные для отдельных отраслей, ведь каждое наблюдение определяется не конкретной отраслью, но парой отраслей, причем под парой понимается неупорядоченное множество из двух элементов.

Тем не менее, именно эта неупорядоченность служит основой для метода, предложенного Diodato, Neffke и O'Clery [23]. Их идея состоит в том, чтобы рассмотреть отдельные модели для каждой отрасли. Набор данных для оценки каждой модели составляют все те наблюдения, в которых данная отрасль i — элемент пары отраслей, определяющих наблюдение. Мы применили изложенный выше подход с перебором всех спецификаций и выделением среди них лучшей по критерию Акаике для каждой отдельной отраслевой группировки из 19. Так как по некоторым наблюдениям из-за нехватки данных или их конфиденциальности отсутствовали оценки интенсивности обмена по матрице Шерера, число наблюдений для некоторых отраслей составляло не 18, но несколько меньше.

Так как существенно взвешивание наблюдений на результаты не повлияло, приводим таблицы только с оценками МНК (таблицы 2 и 3).

Таблица 2

Первые 5 лучших по критерию Акаике моделей для оценки связи трех «маршаллианских» факторов с CL-индексом по данным для всех отраслей в целом

Название параметра / статистики / величины, параметр при которой оценивался	**1**	**2**	**3**	**4**	**5**
Константа	0,0035*** (0,0002)	0,0034*** (0,0002)	0,0035*** (0,0002)	0,0035*** (0,0002)	0,0034*** (0,0002)
Интенсивность обмена по матрице затрат-выпуска			0,0003 (0,0009)		-0,0003 (0,0011)
Сходство требований к рабочей силе	0,0023*** (0,0005)	0,0024*** (0,0005)	0,0023*** (0,0005)	0,0023*** (0,0006)	0,0025*** (0,0005)
Интенсивность обмена по матрице Шерера		0,0013 (0,0010)		0,0006 (0,0022)	0,0015 (0,0012)
Интенсивность обмена по матрице затрат-выпуска × сходство требований к рабочей силе					
Интенсивность обмена по матрице затрат-выпуска × интенсивность обмена по матрице Шерера					
Сходство требований к рабочей силе × интенсивность обмена по матрице Шерера				0,0032 (0,0082)	
R^2	0,1341	0,1432	0,1347	0,1440	0,1438
Приведенный R^2	0,1287	0,1325	0,1239	0,1279	0,1276
p-значение статистики F-критерия	0,0000	0,0000	0,0000	0,0000	0,0000
Логарифм правдоподобия	861,3	862,1	861,3	862,2	862,2
AIC	-1 716,5	-1 716,3	-1 714,7	-1 714,4	-1 714,4
BIC	-1 707,3	-1 703,9	-1 702,3	-1 699,0	-1 698,9
Число наблюдений	163	163	163	163	163

Примечания

1 Составлено по расчетам авторов

2 *** — p-значение меньше 0,01 %; ** — p-значение меньше 1 %; * — p-значение меньше 5 %; . — p-значение меньше 10 %.

Таблица 3

Первые 5 лучших по критерию Акаике моделей для оценки связи трех «маршаллианских» факторов с XCL-индексом по данным для всех отраслей в целом

Название параметра / статистики / величины, параметр при которой оценивался	**1**	**2**	**3**	**4**	**5**
Константа	-0,0011*** (0,0002)	-0,0008*** (0,0001)	-0,0011*** (0,0002)	-0,0011*** (0,0002)	-0,0009*** (0,0002)
Интенсивность обмена по матрице затрат-выпуска	0,0014 (0,0009)	0,0016. (0,0009)	0,0033 (0,0026)	0,0013 (0,0009)	
Сходство требований к рабочей силе	0,0006 (0,0004)		0,0008. (0,0005)	0,0006 (0,0005)	0,0006 (0,0004)
Интенсивность обмена по матрице Шерера	0,0042** (0,0014)	0,0038** (0,0014)	0,0040** (0,0015)	0,0041* (0,0019)	0,0018* (0,0008)
Интенсивность обмена по матрице затрат-выпуска × сходство требований к рабочей силе			-0,0046 (0,0057)		
Интенсивность обмена по матрице затрат-выпуска × интенсивность обмена по матрице Шерера	-0,0092* (0,0039)	-0,0090* (0,0039)	-0,0098* (0,0039)	-0,0093* (0,0041)	
Сходство требований к рабочей силе × интенсивность обмена по матрице Шерера				0,0004 (0,0073)	
R^2	0,0729	0,0597	0,0768	0,0729	0,0376
Приведенный R^2	0,0494	0,0419	0,0474	0,0434	0,0256
p-значение статистики F-критерия	0,0171	0,0202	0,0268	0,0348	0,0466
Логарифм правдоподобия	897,4	896,3	897,8	897,4	894,4
AIC	-1 782,9	-1 782,6	-1 781,5	-1 780,9	-1 780,8
BIC	-1 764,3	-1 767,1	-1 759,9	-1 759,2	-1 768,4
Число наблюдений	163	163	163	163	163

Примечания

1 Составлено по расчетам авторов

2 *** — p-значение меньше 0,01 %; ** — p-значение меньше 1 %; * — p-значение меньше 5 %; . — p-значение меньше 10 %.

Из приведенных выше таблиц следует, что важнее всего большой рынок труда, меньше значит возможность извлекать выгоду из перетока знаний. Близость к поставщикам/покупателям, их разнообразие меньше всего связано с размещением отраслей в тех же районах. Свидетельств взаимодействия разных факторов мало, но одновременное сходство требований к рабочей силе и ценность изобретений друг друга для пары отраслей могут усиливать их сосредоточение. Отрицательная оценка параметра при произведении близости в производственной цепочке и в технологическом пространстве может указывать на

альтернативный источник заимствований новаций: благодаря изобретениям, овеществленным в товарах. Больший XCL-индекс связан прежде всего с возможностью использовать местные перетоки знаний.

Результаты оценок для отдельных видов деятельности следующие. Для разных отраслей различные виды «маршаллианских экстерналий» имеют неодинаковое значение. Размещение организаций сферы обслуживания меньше всего связано с каким-либо маршаллианским механизмом агломерационных выгод (при самом большом числе наблюдений). Самые сильные свидетельства связи сосредоточения со всеми тремя источниками выгод — для транспортного машиностроения (автомобили, самолеты, суда, ракеты и т. п.). Близость к поставщикам/покупателям, их разнообразие привлекает прежде всего автомобилестроение; также производство летательных аппаратов и судов; деревообработку.

Заключение

На основе разбора зарубежных подходов к делимитации методами машинного обучения городских агломераций для применения в российских условиях отобран и приспособлен метод DBSCAN. Проведена апробация метода на данных о центроидах жилых домов в российских регионах с поправкой на оценочное число их жильцов. Выделенные благодаря DBSCAN агломерации дают более оптимальные территориальные единицы для оценки указанных выше микроэкономических функций, так как городские агломерации по определению сосредоточивают внутри себя бóльшую часть связей расположенных в них производителей.

Чтобы сравнить силу трех видов агломерационных эффектов по Маршаллу, был использован подход Эллисона — Глейзера — Керра с уточнениями, связанными с особенностями российских данных. Представленные в отчете оценки — это первый опыт измерения степени сосредоточения российских отраслей по данным обо всех без исключения организациях на конкретную дату (1 января 2020 г.).

Полнота охвата организаций в России привела к выводам, которые вступают в противоречие с общим мнением о сосредоточении производителей в России: наши оценки показывают, что в России пары отраслей обычно рассредоточены относительно друг друга: у большей части отраслей сосредоточение существенно ниже, чем можно было ожидать, исходя из общего размещения этих производств.

Анализ связей между, с одной стороны, сосредоточением пар отраслей, а с другой — интенсивностью обмена товарами и услугами между отраслями, сходством их требований к рабочей силе, а также интенсивностью перетоков знаний между отраслями показал, что в среднем из трех внешних выгод сосредоточения по Маршаллу в России важнее всего большой рынок труда. Что касается случаев с сосредоточением пар отраслей сверх ожидаемого, то было установлено, что такая концентрация связана прежде всего с возможностью использовать местные перетоки знаний.

Полученные оценки связи сосредоточения с различными источниками его внешних выгод подкрепляют те меры государственной политики, которые направлены на поощрение развития крупных городских агломераций с большим и постоянным рынком квалифицированного рабочего труда. При формировании особенно плотных кластеров целесообразно устанавливать для территорий с особым режимом предпринимательства требования к виду деятельности, которые бы согласовывались с оценками интенсивности возможного обмена знаниями между отраслями. Сосредоточение в кластерах эффективнее для тех видов деятельности, у которых возможности перетоков выше.

Одна из трудностей, которая стоит перед оценкой силы агломерационных эффектов — это возможность смешения выгод сосредоточения с последствиями отбора лучших предприятий в крупных городах и кластерах из-за более высокой местной конкуренции. Чтобы оценить опасность такой ошибки, мы применили метод Комба и др. (2012), согласно которому отбор и выгоды сосредоточения по-разному изменяют распределение производительности предприятий. На примере Калининградской области показано, что признаков усечения распределения из-за отбора нет.

Благодарности

Препринт подготовлен на основе материалов научно-исследовательской работы, выполненной в соответствии с государственным заданием РАНХиГС при Президенте Российской Федерации на 2022 год.

Список источников

1. Маршалл А. Принципы политической экономии. Т. 1. Москва: Прогресс, 1983. 415 с.
2. Glaeser E.L., Kallal H.D., Scheinkman J.A., and Shleifer A. Growth in Cities // Journal of Political Economy. 1992. Vol. 100. No. 6. pp. 1126-1152.
3. Jacobs J. The Economy of Cities. New York: Random House, 1969.
4. McCann B.T., Folta T.B. Location matters: where we have been and where we might go in agglomeration research // Journal of management. 2008. Vol. 34. No. 3. pp. 532-565.
5. Ellison G., Glaeser E.L. Geographic concentration in US manufacturing industries: a dartboard approach // Journal of Political Economy. 1997. Vol. 105. No. 5. pp. 889-927.
6. Ellison G., Glaeser E., Kerr W. Data and Empirical Appendix to "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns" // American Economic Association. 2009. URL: https://assets.aeaweb.org/asset-server/articles-attachments/aer/data/june2010/20070331_app.pdf (дата обращения: 08.06.2022).
7. Duranton G., Overman H.G. Testing for Localization Using Micro-Geographic Data // The Review of Economic Studies. 2005. Vol. 72. No. 4. pp. 1077-1106.
8. Nakamura R., Paul C.J.M. Measuring agglomeration // In: Handbook of Regional Growth and Development Theories / Ed. by Capello R., Nijkamp P. Cheltenham: Edward Elgar, 2009. pp. 305–328.
9. Measuring Spatial Concentration // In: Economic Geography : The Integration of Regions and Nations / Ed. by Combes P.P., Mayer T., and Thisse J.F. Princeton, Woodstock: Princeton University Press, 2008. pp. 254–275.
10. Audretsch D.B., Feldman M.P. R&D Spillovers and the Geography of Innovation and Production // The American Economic Review. 1996. Vol. 86. No. 3. pp. 630-640.
11. Rosenthal S.S., Strange W.C. The Determinants of Agglomeration // Journal of Urban Economics. 2001. Vol. 50. No. 2. pp. 191–229.
12. Scherer F. Using Linked Patent and R&D Data to Measure Interindustry Technology Flows // In: R&D, Patents, and Productivity. University of Chicago Press, 1984. pp. 417–464.
13. // OpenStreetMap: [сайт]. URL: https://www.openstreetmap.org (дата обращения: 17.03.2022).
14. Ester M., Kriegel H.P., Sander J., and Xu X. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) // A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 1996. P. 6.

15. Maklin C. DBSCAN Python Example: The Optimal Value For Epsilon (EPS) // Towards Data Science. 2019. URL: https://towardsdatascience.com/machine-learning-clustering-dbscan-determine-the-optimal-value-for-epsilon-eps-python-example-3100091 (дата обращения: 21.03.2022).
16. Rahmah N., Sitanggang S. IOP Conference Series: Earth and Environmental Science // Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. 2016. Vol. 31. P. 5.
17. Gao J. Lecture 4: Density-based Methods URL: https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_density.pdf (дата обращения: 25.05.2022).
18. Пономарев Ю.Ю., Радченко Д.М. Разработка научно-методологических подходов к идентификации фактических границ агломераций в России с учетом пространственного распределения экономической активности : Отчет о НИР. Москва: РАНХиГС, 2020.
19. Combes P.P., Duranton G., Gobillon L., Puga D., and Roux S., "The Productivity Advantages of Large Cities: Distinguishing Agglomeration From Firm Selection," *Econometrica*, Vol. 80, No. 6, 2012. pp. 2543-2594.
20. Yeo I.K., Johnson R.A., "A new family of power transformations to improve normality or symmetry," *Biometrika*, Vol. 87, No. 4, Dec 2000. pp. 954-959.
21. Berk R.A. Statistical Learning from a Regression Perspective. Cham: Springer International Publishing, 2016. 372 pp.
22. Diodato D., Neffke F., and O'Clery N., "Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time," *Journal of Urban Economics*, Vol. 106, Jul 2018. pp. 1-26.
23. Duranton G., Puga D. Micro-foundations of urban agglomeration economies // In: Handbook of Regional and Urban Economics. Elsevier, 2004. pp. 2063-2117.

В СЕРИИ ПРЕПРИНТОВ РАНХиГС РАССМАТРИВАЮТСЯ ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ ПОДХОДЫ К СОЗДАНИЮ, АКТИВНОМУ ИСПОЛЬЗОВАНИЮ ВОЗМОЖНОСТЕЙ ИННОВАЦИЙ В РАЗЛИЧНЫХ СФЕРАХ ЭКОНОМИКИ КАК КЛЮЧЕВОГО УСЛОВИЯ ЭФФЕКТИВНОГО УПРАВЛЕНИЯ

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ