ческим обоснованием, особенно для больших данных, приведет к научным результатам, основанным на подходящих подходах. В конечном счете, только сбалансированное взаимодействие всех вовлеченных наук приведет к успешным решениям в науке о данных.

Библиографический список

1. Press, G.: A Very Short History of Data Science // Forbes. May 28, 2013. – URL: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#5ec53e6055cf

2. Tukey J. W. Exploratory Data Analysis. Addison – Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney, 1977. – 688 с.

3. Piateski G., Frawley W. Knowledge Discovery in Databases. – MIT Press, Cambridge, – 1991. – 540 с.

4. Cao L. Data science: a comprehensive overview //ACM Computing Surveys (CSUR). – 2017. – Т. 50. – №. 3. – С. 1-42.

5. Donoho D. 50 years of data science //Journal of Computational and Graphical Statistics. – 2017. – Т. 26. – №. 4. – С. 745-766.

6. Wu J. Statistics = data science? – 1997. – URL: http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

7. Van Dyk D., Fuentes M., Jordan M. I., Newton M., Ray B. K., Lang D. T., Wickham H. ASA Statement on the Role of Statistics in Data Science //Amstat news. – 2015. – Т. 460. – №. 9. – 24 с.

8. Brown M. S. Data mining for dummies. – John Wiley & Sons, London. – 2014. – 410 с.

УДК 519.24; 004.67

Сравнительный анализ методов оценки причинного эффекта: оценка вклада элементов интенсификации технологии в урожайность яровой пшеницы

К.О. Тарасов, Е.В. Понькина

АлтГУ, г. Барнаул

Аннотация. Задача оценки причинных эффектов представляет собой сравнение состояния объекта с учетом и без учета вмешательства и оценки ожидаемой величины полученных различий целевого признака. В работе рассматривается сравнительный анализ методов оценки причинных эффектов, включая тесты попарных сравнений, линейные регрессионные модели и метод псевдорандомизации

(*Propensity Score Matching*). Прикладная задача исследования заключается в оценке эффекта технологии возделывания яровой пшеницы на ее урожайность в условиях Кулундинской степи Алтайского края.

Ключевые слова: причинный эффект, парное сравнение средних, Propensity Score Matching, урожайность пшеницы, интенсивная технология, Алтайский край, Россия.

Введение. Одним из разновидностей анализа эффектов вмешательства (целенаправленного воздействия) в производственно-хозяйственную систему является причинный вывод (*Causal Inference*). Анализ причинности помогает получить заключение о каких-либо гипотезах путем нахождения различий между фактами и гипотетическими ситуациями (контрфактуалами). Фундаментальной основой исследования причинных эффектов в прикладных научных исследованиях является теория о потенциальных исходах (*Potential Outcome Theory*), основанная на результатах Дж. Неймана и Д. Рубина [3]. Центральная проблема теории потенциальных исходов – проблема количественной оценки причинного эффекта воздействия на состояние объекта.

Причинный эффект (*Causal effect*) – это разность между двумя потенциальными исходами объекта Y(1) и Y(0), где при прочих равных условиях Y(1) – оценка результирующего признака объекта в результате вмешательства, Y(0) – оценка результирующего признака без учета вмешательства [4].

Для решения этой проблемы Дж. Нейман и Д. Рубин [5] предложили модель Рубина-Неймана, которая позволяет осуществлять псевдорандомизацию данных, с целью приблизить условия исследования на базе обсервационных данных к условиям экспериментального исследования, тем самым уменьшить смещение в оценке причинного эффекта и приблизить полученную оценку к истинному причинному эффекту.

Целью исследования является сравнительный анализ методов оценки причинных эффектов на примере оценки вклада элементов интенсификации технологии возделывания яровой пшеницы на территории Кулундинской степи в границах Алтайского края в засушливых условиях 2012 года.

Методы анализа причинности. Для анализа причинности в работе используются методы попарного сравнения средних – тест t-критерия Стьюдента (t-тест) [8] и тест U-критерия Манна-Уитни-Вилкоксона [9], логистическая регрессия и метод сопоставления оценок склонностей (*Propensity Score Matching – PSM*). Метод сопоставления оценок склонностей (PSM) был разработан Дональдом Рубином совместно с

Полом Розенбаумом для того, чтобы уменьшить предвзятость, связанную с переменными-конфаундерами, которая может быть обнаружена при расчете эффекта воздействия. Конфаундеры (*Confounder*) – это переменные, которые могут быть связаны как с изучаемым фактором воздействия, так и с выводом [1]. Их действие создает различия между рассчитанным и фактическим эффектом воздействия.

Можно выделить шесть основных шагов при реализации метода PSM:

1. Сбор и анализ данных.
2. Отбор ковариантов для использования в модели соответствия.
3. Выбор модели для расчета оценок склонности объектов и расчет этих оценок.
4. Выбор метода анализа соответствий (Matching) и поиск соответствий.
5. Анализ балансировки данных.
6. Расчет причинного эффекта.

Данные и территория исследования. Кулундинская степная зона (далее – Кулунда) расположена в юго-восточной части Западной Сибири и простирается от центра до юга Алтайского края. Климат в Кулунде континентальный, многолетняя средняя температура – минус 18 °С в самый холодный месяц (январь) и +19 °С в самый теплый месяц (июль) [7]. Число дней с температурой выше +5°С колеблется от 153 в северной части до 173 в южной части Кулунды.

Исходные данные представляют собой результаты опроса (интервью) руководителей сельскохозяйственных предприятий в комбинации со статистическими данными официальной статистики мониторинга деятельности предприятий [2]. Общее количество наблюдений – 196, из них 111 предприятий относятся к категории обществ с ограниченной ответственностью. Средняя площадь земель сельскохозяйственного назначения на предприятии составила 9333 га, а посевная площадь 7150 га. Данные включают индикаторы продуктивности яровой пшеницы в 2012 г. и в среднем за 2008-2012 гг., показатели масштаба деятельности (площадь посева, площадь земель сельхозназначения, количество рабочих), показатели технологии возделывания пшеницы (применение чередования культур в севообороте, внесение удобрений и средств защиты растений, доля паров в структуре посева), индикаторы специализации производства (доля зерновых и зернобобовых культур в структуре посева, доля выручки от реализации продукции растениеводства), характеристики менеджера предприятия (возраст, образова-

ние), климатические параметры возделывания пшеницы (сумма осадков за вегетационный период).

В целом предприятия классифицируются на пары групп (*тестовая группа* – объекты испытывающие вмешательство и *контрольная группа* – объекты не испытывающие вмешательство) по бинарным признакам в следующих вариантах:

– применение удобрений (да/нет);

– применение средств защиты растений (да/нет);

– применение удобрений и средств защиты растений (да&да/нет&нет).

Оценка причинного эффекта в виде прироста урожайности яровой пшеницы осуществляется посредством сравнения урожайности яровой пшеницы в тестовой и контрольной группах.

Предварительный анализ данных показал, что только 24% фермеров вносило удобрения в технологии возделывания пшеницы, а 69% применяли средства защиты растений. В 2012 году в следствие засухи наблюдалась низкая средняя урожайность пшеницы – 6,01 ц/га (максимум достигал 20 ц/га), при средней урожайности за 2008-2012 гг. – 9,36 ц/га (максимум 26 ц/га). Значительно более низкая продуктивность пшеницы в 2012 году объясняется малым количеством осадков – 86 мм в среднем по территории в период с апреля по август.

Результаты. Результаты оценки причинных эффектов на основе различных методов подтвердили, что применение двух компонент технологии (средств защиты растений и удобрений) способствует увеличению урожайности яровой пшеницы с каждого гектара как минимум на 3,12 центнера. Это по приблизительным оценкам принесло бы дополнительно 2652 рублей дополнительной прибыли с каждого гектара пашни в ценах 2012 г., при средней цене реализации за тонну зерна в 8500 рублей [6], а средних затратах на СЗР – 280 руб./га, на удобрения – 328 руб./га. Прирост чистой прибыли при применении только СЗР составил бы 2044 руб./га. В целом оценки причинных эффектов варьируются от 0,39 до 3,23 ц/га в случае внесения удобрений, применение СЗР дает эффект варьируемый от 0,94 до 3,52 ц/га. Интенсификация технологии и применении обоих компонент дает эффект от 1,34 до 5,68 ц/га. Таким образом, можно констатировать позитивный эффект интенсификации, обнаруженный по данным практики возделывания пшеницы в Алтайском крае.

Таблица 1 – Результаты оценки причинных эффектов интенсификации технологии возделывания пшеницы, Алтайский край

	Уровни интенсификации технологии возделывания пшеницы:		
	Внесение удобрений	*Применение СЗР*	*Применение СЗР и удобрений*
Парное сравнение средних:			
t-тест с различными дисперсиями	1,85	2,11	3,47
Тест Манна-Уитни-Вилкоксона	1,60	2,00	3,20
Регрессионный анализ:			
Однофакторная регрессия	1,85	2,11	3,47
Многофакторная регрессия	1,50	2,06	3,12
Propensity Score Matching:			
Метод полного соответствия	1,50	2,06	3,12
Метод генетического соответствия	1,81	2,30	3,71

Источник: Вычисления авторов.

Заключение. Преимущества метода PSM относительно других методов заключаются, в первую очередь, в применении псевдо-рандомизации данных, что позволяет считать полученную оценку причинного эффекта приближенной к истинному эффекту. В целом оценки по методу PSM показывают несущественные различия с результатами, полученными по другим методам (различие менее 0,5 ц/га считаем несущественным). Метод PSM позволяет проводить приближенные к экспериментальным условиям исследования на больших массивах данных, а также с данными мониторинга производственно-хозяйственной деятельности, в случаях когда провести экспериментальное исследование с полностью рандомизированной выборкой невозможно или крайне сложно. Главное преимущество метода PSM заключается в балансировке тестовой и контрольной выборок данных по набору контрольных признаков (например, параметры масштаба деятельности, организационно-правовая форма, специализация производства), обеспечивающей корректное сравнением объектов тестовой и контрольной групп.

Библиографический список

1. Clinical Pharmacy Education, Practice and Research. – Elsevier, 2019.
2. Revealing the determinants of wheat yields in the Siberian breadbasket of Russia with Bayesian networks / A.V. Prishchepov [и др.] // Land Use Policy. – 2019. – Т. 80. – С. 21-31.
3. Rubin D.B. Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies / D.B. Rubin. – С. 140.
4. Yamamoto T. Statistical Models for Causal Analysis / T. Yamamoto. – С. 81.
5. Imbens G.W. Rubin Causal Model / G.W. Imbens, D.B. Rubin // The New Palgrave Dictionary of Economics / ред. Palgrave Macmillan. – London: Palgrave Macmillan UK, 2008. – С. 1-10.
6. Мониторинг цен на пшеницу организаций 01.10.2012: АгроНовости Ассет [Электронный ресурс]. – URL: https://agro-bursa.ru/prices/wheat/01-10-2012/ (дата обращения: 06.06.2020).
7. AgroAtlas - Главная [Электронный ресурс]. – URL: http://www.agroatlas.ru/ru/index.html (дата обращения: 03.06.2020).
8. Kalpić D. Student's t-Tests / D. Kalpić, N. Hlupić, M. Lovrić // International Encyclopedia of Statistical Science / ред. M. Lovric. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. – С. 1559-1563.
9. Neuhäuser M. Wilcoxon–Mann–Whitney Test / M. Neuhäuser // International Encyclopedia of Statistical Science / ред. M. Lovric. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. – С. 1656-1658.

УДК 004

Реализация системы визуального представления статистической информации по работе УВД Алтайского края

А.В. Турчановская, О.Н. Половикова
АлтГУ, г. Барнаул

Метод визуализации – графическое представление, упрощающее анализ теоретических и статистических данных, процесс восприятия и осмысления информации, формирование новых навыков и умений. Использование изображений делает процесс изучения объектов более наглядным. Основная идея методов визуализации заключается в представлении человеку-эксперту большого объема данных в дос-