УДК 004.852:519.218.82:338.43(575.2)
DOI: 10.36979/1694-500X-2026-26-8-74-82

ВАРИАЦИОННЫЙ АВТОЭНКОДЕР ДЛЯ ВОССТАНОВЛЕНИЯ ПРОПУСКОВ В ПАНЕЛИ АГРОЭКОНОМИЧЕСКИХ ДАННЫХ КЫРГЫЗСКОЙ РЕСПУБЛИКИ: АРХИТЕКТУРА, АУГМЕНТАЦИЯ И ВАЛИДАЦИЯ МЕТОДОМ SELF-MASKING

Ж.Б. Мамадалиева, А.Д. Картанова, Ж.Р. Сарыпбекова

Аннотация. Статья продолжает работу, в которой описано формирование сбалансированной панели агроэкономических данных Кыргызской Республики методом линейной интерполяции и множественного вменения цепными уравнениями (MICE) на базе алгоритма Random Forest. В настоящем исследовании рассматриваются альтернативный и дополняющий подходы к иной, более узкой задаче – подготовке данных непосредственно для обучения глубокой прогностической нейронной сети: маскированный вариационный автоэнкодер (VAE), генеративная модель, обучаемая напрямую на неполных наблюдениях. Статья носит методологический характер: подробно описаны архитектура энкодера и декодера с явным маскированием пропусков, агрономически взвешенная функция потерь ELBO, пятикомпонентная схема аугментации обучающей выборки, метод оценки неопределённости Monte Carlo Dropout и процедура self-masking валидации, позволяющая объективно оценить качество импутации в отсутствие истинных значений для реально пропущенных наблюдений, в сравнении с тремя эталонными методами – средним k ближайших соседей (KNN) и MICE (IterativeImputer). Количественные результаты применения описанной методологии на конкретном запуске пайплайна приведены в отдельной статье того же цикла исследований.

Ключевые слова: импутация пропущенных данных; вариационный автоэнкодер; MICE; IterativeImputer; self-masking; Monte Carlo Dropout; аугментация данных; агроэкономическая панель; Кыргызская Республика.

КЫРГЫЗ РЕСПУБЛИКАСЫНЫН АГРОЭКОНОМИКАЛЫК БЕРИЛИШТЕР ПАНЕЛИНДЕГИ БОШ ОРУНДАРДЫ ТОЛТУРУУ ҮЧҮН ВАРИАЦИЯЛЫК АВТОЭНКОДЕР: АРХИТЕКТУРАСЫ, АУГМЕНТАЦИЯ ЖАНА SELF-MASKING ЫКМАСЫ МЕНЕН ВАЛИДАЦИЯ

Ж.Б. Мамадалиева, А.Д. Картанова, Ж.Р. Сарыпбекова

Аннотация. Бул макала эмгектин уландысы болуп саналат. Анда Кыргыз Республикасынын агроэкономикалык берилиштеринин Random Forest алгоритминин негизинде, сызыктуу интерполяция жана чынжырлуу теңдемелер аркылуу көптүк импутация (MICE) ыкмасы менен түзүлгөн тең салмактуу панели сүрөттөлгөн. Учурдагы изилдөөдө терең прогноздоочу нейрондук тармакты окутуу үчүн берилиштерди даярдоонун альтернативдик жана толуктоочу мамилеси каралат: бул – толук эмес байкоолордун негизинде түздөн-түз окутулган генеративдик модель болгон маскировкаланган вариациялык автоэнкодер (VAE). Макала усулдук мүнөзгө ээ. Анда төмөнкү элементтер кеңири сүрөттөлгөн: берилиштердеги боштуктарды (же жок берилиштерди) ачык маскировкалоо менен түзүлгөн энкодер жана декодердин архитектурасы; агрономиялык жактан салмакталган ELBO жоготуу функциясы; окутуу тандамын аугментациялоонун беш компоненттүү чийими; Monte Carlo Dropout ыкмасы менен белгисиздикти баалоо жана чын мааниси белгисиз болгон накта жоголгон байкоолор үчүн импутациянын сапатын объективдүү баалоого мүмкүндүк берген self-masking валидация жол-жобосу. Аталган ыкмалар үч эталондук усул - орточо маани, k эң жакын кошуналар (KNN) жана MICE менен салыштырылып изилденген. Сүрөттөлгөн усулдукту конкреттүү пайплайнды иштетүүдө колдонуунун сандык натыйжалары ошол эле изилдөө циклинин башка бир макаласында келтирилген.

Түйүндүү сөздөр: жетишпеген маалыматтарды импутациялоо; вариациялык автоэнкодер; MICE; IterativeImputer; self-masking; Monte Carlo Dropout; берилиштерди аугментациялоо; агроэкономикалык панель; Кыргыз Республикасы.

VARIATIONAL AUTOENCODER FOR MISSING DATA IMPUTATION IN THE AGROECONOMIC DATA PANEL OF THE KYRGYZ REPUBLIC: ARCHITECTURE, AUGMENTATION, AND SELF-MASKING VALIDATION

Zh.B. Mamadalieva, A.D. Kartanova, Zh.R. Sarypbekova

Abstract. This article continues the work, which described the construction of a balanced panel of agroeconomic data for the Kyrgyz Republic using linear interpolation and multiple imputation by chained equations (MICE) based on the Random Forest algorithm. The present study considers an alternative and complementary approach to a different, more narrowly defined task: preparing data directly for training a deep predictive neural network, namely, a masked variational autoencoder (VAE), a generative model trained directly on incomplete observations. The article is methodological in nature and describes in detail the encoder-decoder architecture with explicit masking of missing values, an agronomically weighted ELBO loss function, a five-component training-set augmentation scheme, the Monte Carlo Dropout method for uncertainty estimation, and a self-masking validation procedure that enables an objective assessment of imputation quality in the absence of true values for genuinely missing observations. The proposed approach is benchmarked against three baseline methods: mean imputation, k-nearest neighbors (KNN), and MICE (IterativeImputer). Quantitative results obtained from applying the described methodology in a specific pipeline run are presented in a separate article within the same research series.

Keywords: missing data imputation; variational autoencoder; MICE; IterativeImputer; self-masking; Monte Carlo Dropout; data augmentation; agroeconomic panel; Kyrgyz Republic.

Введение. В работе [1] показано, что панель агроэкономических данных Кыргызской Республики (девять регионов, десять групп культур, 1991–2025 гг. [2], источник – Национальный статистический комитет КР [3]) содержит пропуски трёх принципиально разных типов, требующих разных методов обработки: структурные нули, обусловленные природными ограничениями возделывания культур в конкретных регионах; временны́е лакуны, связанные со сменой методологии учёта; и систематическое несовпадение хронологических горизонтов отдельных показателей. Для устранения временны́х лакун в [1] предложен и обоснован конвейер из линейной интерполяции и MICE на основе Random Forest [4–6], ориентированный на задачи эконометрического анализа панели в целом.

Настоящая статья не заменяет этот конвейер, а исследует альтернативу для смежной, но иной по масштабу задачи – подготовку данных непосредственно перед обучением глубокой прогностической модели урожайности, где помимо точности восстановления значения важны также устойчивость метода при очень малом объёме данных (315 наблюдений «регион – год» – девять регионов на протяжении 1991–2025 гг. – вместо тысяч, используемых в [1]), явный учёт агрономической недопустимости отдельных комбинаций «регион – культура» и количественная оценка неопределённости каждого восстановленного значения. Для решения этой задачи используется маскированный вариационный автоэнкодер (VAE) – генеративная нейросетевая модель, восстанавливающая пропуски совместно по всем признакам, а не по одному показателю за раз.

Эталонные методы импутации. Импутацией называют замену отсутствующих значений в наборе данных правдоподобными оценками, вычисленными на основе наблюдаемой части данных. Простейший вариант – заполнение каждого пропуска средним (или медианным) значением соответствующего столбца – служит нижней границей качества: он занижает дисперсию данных и разрушает связи между признаками (например, между урожайностью и суммой активных температур), поэтому используется здесь исключительно как контрольный ориентир, а не как содержательный метод.

Метод k ближайших соседей (K-Nearest Neighbors, KNN) восстанавливает пропущенное значение как среднее по k (в настоящем исследовании $k = 5$) наиболее похожим наблюдениям, где похожесть измеряется расстоянием в пространстве остальных, наблюдаемых признаков [7, 8]. Учёт локального сходства делает KNN точнее среднего, однако при большом числе признаков (109 в используемой панели) понятие «похожести» становится менее надёжным.

MICE (Multivariate Imputation by Chained Equations) – метод множественного вменения цепными уравнениями, формализованный ван Бюреном и Гротхёйс-Аудсхорн [4], подробно рассмотренный

в современных обзорах [9, 10] и лежащий в основе конвейера, описанного в [1]: пропуски восстанавливаются не за один проход, а итеративно – для каждого признака строится регрессия на основе всех остальных признаков, и цикл повторяется до стабилизации оценок. В программной среде Python метод реализован в виде класса IterativeImputer библиотеки scikit-learn [11]. В [1] MICE использовался с алгоритмом Random Forest [6] в качестве регрессионной модели на каждом шаге цепочки и восстанавливал панель из тысяч наблюдений для целей эконометрического анализа; здесь MICE выступает эталоном сравнения для существенно более узкой задачи – импутации выборки из 315 наблюдений и 109 признаков непосредственно перед обучением нейросетевой прогностической модели, с настройками, приближёнными к линейной регрессии на каждом шаге.

Вариационный автоэнкодер (Variational Autoencoder, VAE) [12, 13], получивший дальнейшее развитие в задачах импутации неполных данных [14], – разновидность автоэнкодера [15], в которой скрытое представление данных параметризуется не единственной фиксированной точкой, а параметрами вероятностного распределения (нормального, со средним μ и разбросом σ), из которого затем производится сэмплирование. В отличие от KNN и MICE, VAE не ищет похожие наблюдения и не строит отдельную регрессию на каждый признак, а обучает единую модель совместного распределения сразу всех 109 признаков, что потенциально позволяет учитывать более сложные, нелинейные взаимосвязи климата, экономики и урожайности – ценой того, что такой модели требуется больше обучающих данных, чем доступно в исходной выборке.

Архитектура маскированного энкодера и декодера. Обычный автоэнкодер не способен отличить «настоящий ноль» (например, действительно нулевой импорт удобрений в конкретном году) от «значение неизвестно». Для устранения этой неоднозначности в архитектуру введена маска наблюдаемости M – бинарный массив той же формы, что и данные, где 1 означает «значение известно», а 0 – «значение отсутствует». На вход энкодера подаётся конкатенация двух частей: произведения $x \odot M$ (значения признаков, умноженные поэлементно на маску) и самой маски M, так что итоговая размерность входа равна не F, а 2F.

Энкодер и декодер устроены симметрично (рисунок 1): три последовательных полносвязных слоя сжимают размерность признаков с 2F до 256, 128 и 64 нейронов (энкодер) и зеркально расширяют её обратно (64→128→256→F, декодер). После каждого слоя применяется нормализация LayerNorm [16], активация ELU [17] и Dropout с интенсивностью 0,1. Активация ELU выбрана вместо более распространённой ReLU потому, что устойчивая к выбросам нормализация RobustScaler не ограничивает значения диапазоном (0, 1): отрицательные значения – полноценная часть входных данных, и активация, обнуляющая их, приводила бы к потере информации. Между энкодером и декодером расположено латентное пространство размерности 32; размерности слоёв 256, 128, 64 и 32 не зависят от числа признаков F и фиксированы архитектурой. Для используемой панели F = 109 (девять регионов, семь групп признаков: урожайность – 15 показателей, климатические – 17, экономические – 56, площадь посева – 15, техника – 4, лаги цены продукции – 6, служебные флаги – 3), то есть вход энкодера имеет размерность 2F = 218; именно параметры распределения в латентном пространстве (μ и логарифм дисперсии) выучивает энкодер.

Агрономически взвешенная функция потерь ELBO.

$$L_{VAE} = \Sigma[(x_{recon} - x_{true})^2 . M . \left(1 + A_{agro}\right] / \Sigma\left[M . \left(1 + A_{agro}\right)\right] + \beta . KL(q\left(zVx\right) | |\rho\left(z\right)), \beta = 0,1 .$$

Стандартная для вариационных автоэнкодеров функция потерь ELBO (Evidence Lower BOund) состоит из члена реконструкции, измеряющего точность восстановления наблюдаемых значений, и расхождения Кульбака – Лейблера (KL-дивергенции), не позволяющего энкодеру выучить неструктурированное скрытое пространство, пригодное лишь для запоминания обучающих примеров [12, 18]. В настоящей задаче стандартный ELBO модифицирован двумя способами. Во-первых, член реконструкции считается только по фактически наблюдаемым позициям (домножается на маску M) – модель не наказывается за несовпадение с уже отсутствующими значениями. Помимо этого, вклад

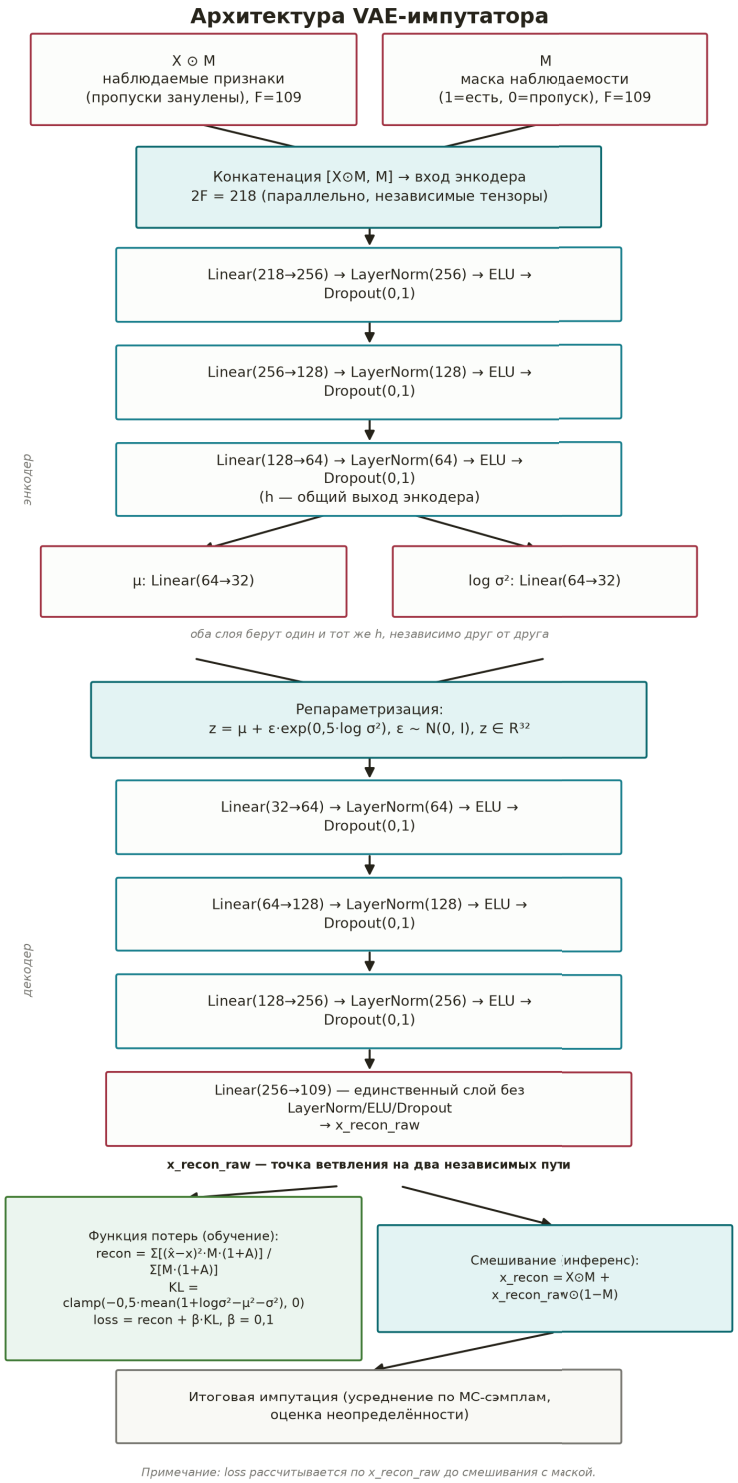


Рисунок 1 – Архитектура маскированного вариационного автоэнкодера: вход $x \odot M$ и маска M, энкодер 2F→256→128→64, латентное пространство (j , $\log\Gamma I$) размерности 32, декодер 64→128→256→F. Для используемой панели F = 109, то есть 2F = 218

каждого наблюдения взвешивается ещё и агрономической маской достоверности: наблюдениям, относящимся к агрономически допустимым сочетаниям «регион – культура», присваивается повышенный вес (множитель $1 + A_{agro}$) в сравнении со структурными нулями.

Использование коэффициента $\beta = 0,1$ вместо единицы, как в классическом ELBO, смещает приоритет в сторону точности реконструкции в ущерб структурированности скрытого пространства. Такой приём в литературе получил отдельное название $\beta - VAE$ [18] и оправдан тем, что применительно к задаче импутации важно именно точное восстановление значений, а не интерпретируемость латентных факторов.

Аугментация обучающей выборки. Главное ограничение задачи – скромный объём исходных данных: после отбраковки записей без указания региона или года остаётся всего 315 наблюдений «регион – год» при 109 признаках (девять регионов на протяжении 35 лет, 1991–2025 годы), в то время как в литературе VAE обычно обучают на выборках в десятки и сотни тысяч примеров. Чтобы сгладить этот разрыв, не нарушая при этом содержательной структуры данных, был выстроен набор из пяти приёмов аугментации (рисунок 2), применяемых совместно; исходные 315 строк остаются в обучающем пуле без изменений, а сам пул пополняется случайной выборкой сгенерированных кандидатов до 2500 наблюдений (прирост примерно в 7,9 раза).

Гауссов шум добавляется к каждому наблюдаемому значению из нормального распределения (σ от 0,01 до 0,03), с амплитудой, увеличенной в 2–3 раза для записей 1991–1995 годов и отдельных



Рисунок 2 – Пятикомпонентная схема аугментации обучающей выборки, который расширяет пул наблюдений с 315 до 2500 при контроле искажения корреляционной структуры

кризисных лет (1992, 1993, 2010, 2020), что заранее учитывает пониженную надёжность данных этого периода. Целевые климатические сценарии вносят вариации только в климатические переменные (коэффициент Селянинова [19], температурные и осадковые аномалии, число морозных дней и др.) с индивидуально подобранной для каждой переменной амплитудой. Региональный MixUp [20] смешивает пары агроклиматически похожих регионов (например, Баткен и Джалал-Абад) в пределах одного года по формуле $X_{=2} = \gamma . X_a + (1-\gamma) . X_b$, где $\lambda \sim Beta(0,2;0,2)$ – распределение, чаще дающее значения, близкие к 0 или 1, что реалистичнее отражает поведение соседних, но не идентичных территорий. Bootstrap [21, 22] формирует новые выборки повторным извлечением наблюдений с добавлением слабого шума ($\sigma = 0,005$), увеличивая эффективный объём выборки конкретного региона. Случайное дополнительное маскирование – аналог denoising autoencoder и self-supervised обучения [23] – искусственно скрывает 15 % уже наблюдаемых значений (30 % для 1991–1995 гг.), заставляя модель восстанавливать их так же, как настоящие пропуски.

Правильность аугментации проверяется автоматически: по отобранным ключевым признакам (урожайность основных культур, климатические индексы) вычисляется среднее абсолютное отклонение корреляционной структуры данных до и после добавления искусственных примеров, и если оно превышает порог 0,1, выводится предупреждение – эта проверка страхует от ситуации, когда аугментация неправдоподобно расширяет данные, а искажает содержательные закономерности. Значение отклонения для конкретного запуска пайплайна и соответствующие коррелограммы приведены в статье [24].

Оценка неопределённости: Monte Carlo Dropout. Значение, восстановленное моделью, остаётся лишь точечной оценкой и не гарантирует совпадения с истинным – для реальных пропусков оно принципиально неизвестно. Dropout, используемый при обучении для регуляризации [25], в обычной практике на этапе инференса отключают. Monte Carlo Dropout (MC Dropout) [26], напротив, оставляет Dropout включённым и при применении модели: одно и то же наблюдение прогоняется через сеть 100 раз, причём каждый раз случайным образом отключается свой набор нейронов. По совокупности таких проходов строится эмпирическое распределение возможных значений, из которого вычисляются среднее (это и есть восстановленное значение), стандартное отклонение и границы 95%-го интервала (2,5-й и 97,5-й процентили) – недорогой в вычислительном плане способ приближённого байесовского вывода. Значения с разбросом свыше двух стандартных отклонений помечаются как ненадёжные и без дополнительной проверки не используются. Диагностика распределения прогнозов на отдельных наблюдениях приведена в статье [24].

Self-masking валидация: методология сравнения. Любая процедура импутации сталкивается с одной и той же методологической проблемой: значение, отсутствующее в данных, по определению отсутствует, и сравнить восстановленное значение с истинным для реальных пропусков напрямую нельзя. Self-masking валидация [23, 27] обходит эту проблему иначе: из заведомо известных данных случайно «стирается» часть значений, которые затем восстанавливаются тестируемым методом и сопоставляются с истинными – ведь истинный ответ всё это время был известен исследователю.

В данном исследовании self-masking (рисунок 3) проводится на трёх контрольных культурах – пшенице, хлопке и сахарной свёкле, – выбранных как показательные примеры соответственно стабильной, структурно нестационарной и статистически «шумной» (склонной к выбросам) динамики. При фиксированном seed = 42 (для воспроизводимости) из полностью наблюдаемых значений урожайности этих культур скрывается 20 % наблюдений; их восстанавливают все четыре сравниваемых метода (VAE, среднее, KNN-5, MICE/IterativeImputer), после чего для каждого метода рассчитываются RMSE, MAE, MAPE, R^2 и корреляция Пирсона. Отдельно для сахарной свёклы, чьи наблюдаемые данные содержат единичные аномально высокие значения, подсчитывается число импутированных значений, превышающих физически неправдоподобный порог 500 ц/га. Процедура проводится

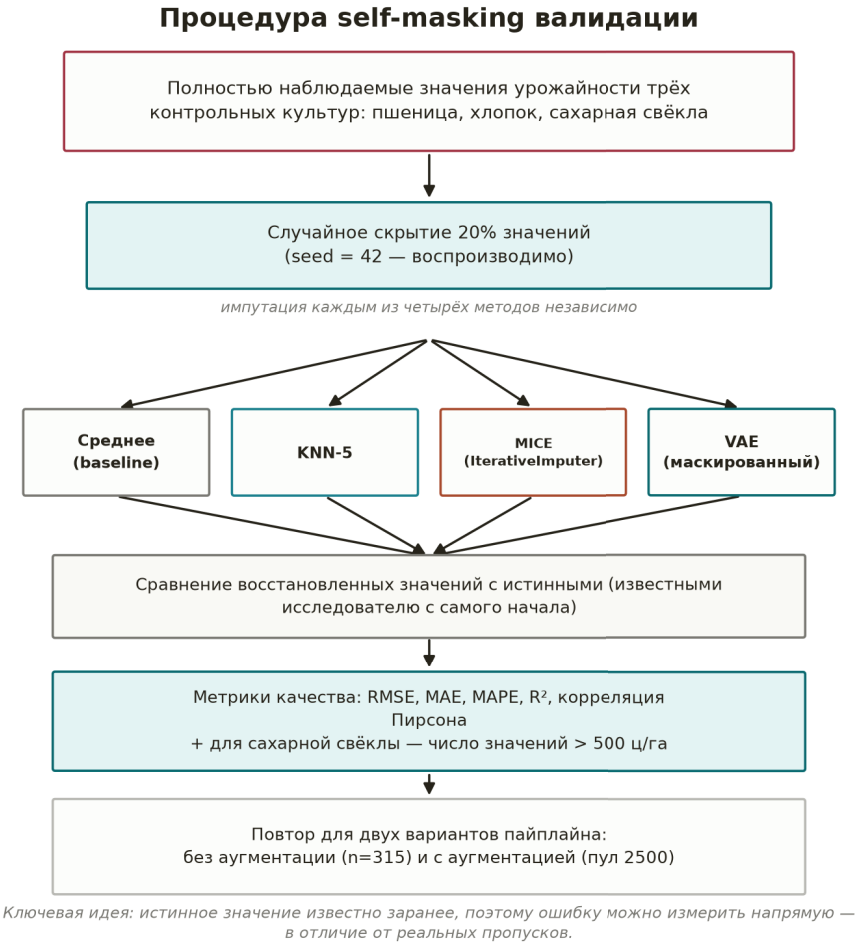


Рисунок 3 – Схема процедуры self-masking валидации: искусственное скрытие 20 % реально наблюдаемых значений, независимая импутация четырьмя методами, сравнение с истинными значениями и расчёт метрик качества

дважды – без аугментации и с полной пятикомпонентной аугментацией, – что позволяет напрямую оценить эффект аугментации на итоговое качество импутации.

Обсуждение. Сравнение VAE с MICE в рамках этой методологии стоит рассматривать не как заранее предрешённое соревнование «какой метод лучше», а как открытый эмпирический вопрос, ответ на который зависит от объёма и структуры конкретной панели. С теоретической точки зрения у VAE есть потенциальное преимущество: восстановление всех показателей происходит совместно, а не последовательно по одному признаку, агрономические ограничения учитываются напрямую через взвешенную функцию потерь, а неопределённость оценивается количественно через MC Dropout – то, чего детерминированная реализация IterativeImputer в MICE предложить не может; сравнительный анализ подобных подходов на данных официальной статистики представлен в [28]. При этом крайне малый объём исходной панели (порядка нескольких сотен наблюдений «регион – год») – на порядки меньше типичных объёмов, на которых в литературе успешно обучаются генеративные нейросетевые модели, – теоретически способен свести это преимущество на нет: MICE, опирающийся на существенно более простую, почти линейную индуктивную модель на каждом шаге цепочки, может обобщаться

устойчивее именно на малых выборках, тогда как многослойная сеть VAE рискует недообучиться или переобучиться даже после аугментации обучающего пула.

Столь же неоднозначен и ожидаемый эффект аугментации: с одной стороны, пять ранее описанных приёмов призваны компенсировать нехватку данных и сделать обучение устойчивее; с другой – синтетические примеры по построению отличаются от распределения реальных наблюдений, и это отличие теоретически может как пойти на пользу (регуляризация, устойчивость к шуму), так и навредить, сместив обучающее распределение относительно контрольных self-masking значений; практические аспекты выбора и реализации методов импутации на реальных данных обсуждаются также в [29–34].

Дать ответ на эти вопросы логическим выводом из архитектуры модели невозможно – он зависит от эмпирических свойств конкретной выборки и требует прямой проверки на self-masking валидации, что и составляет предмет отдельного эмпирического анализа в рамках настоящего цикла исследований.

Заключение. В статье описана методология импутации пропущенных значений в панели агроэкономических данных Кыргызской Республики методом маскированного вариационного автоэнкодера – как самостоятельное исследование, продолжающее и дополняющее конвейер на основе линейной интерполяции и MICE с Random Forest, описанный в [1]. Даны архитектура энкодера и декодера с явным маскированием, вывод и обоснование агрономически взвешенной функции потерь ELBO, пятикомпонентная схема аугментации обучающей выборки, метод MC Dropout для оценки неопределённости и процедура self-masking валидации – необходимый инструмент объективной проверки качества импутации в условиях, когда истинные значения реально пропущенных наблюдений принципиально недоступны. Количественная проверка описанной методологии на конкретном запуске пайплайна приведена и подробно обсуждается в отдельной статье [24] того же цикла исследований.

Поступила: 03.08.2026; рецензирована: 17.08.2026; принята: 19.08.2026.

Литература

1. Формирование сбалансированной панели агроэкономических данных Кыргызской Республики: источники, структура и методы восстановления пропусков // Настоящий цикл исследований, статья в печати. Известия КГТУ им. И. Раззакова. 2026. № 3 (79).
2. *Картанова А.Д.* Создание базы верифицированных данных о сельхозпроизводителях: шаг к цифровой трансформации АПК / А.Д. Картанова, Ж.Р. Сарыпбекова, Ж.Б. Мамадалиева, З.А. Сыдыкова // Вестник Международного Университета Кыргызстана. 2025. № 2-1 (59). С. 71–81. URL: https://www.elibrary.ru/item.asp?id=89032482 (дата обращения: 13.07.2026).
3. Национальный статистический комитет Кыргызской Республики. Статистические сборники по сельскому хозяйству, 1991–2025. Бишкек. URL: http://www.stat.kg (дата обращения: 13.07.2026).
4. *van Buuren S., Groothuis-Oudshoorn K*. Mice: Multivariate Imputation by Chained Equations in R // Journal of Statistical Software. 2011. Vol. 45. No. 3. P. 1–67.
5. *Rubin D.B.* Multiple Imputation for Nonresponse in Surveys. New York: Wiley, 1987. 258 p.
6. *Breiman L.* Random Forests // Machine Learning. 2001. Vol. 45. P. 5–32.
7. *Troyanskaya O., Cantor M., Sherlock G. et al*. Missing value estimation methods for DNA microarrays // Bioinformatics. 2001. Vol. 17. No. 6. P. 520–525.
8. *Cover T., Hart P.* Nearest neighbor pattern classification // IEEE Transactions on Information Theory. 1967. Vol. 13. № 1. P. 21–27.
9. *Rezende D.J., Mohamed S.* Variational Inference with Normalizing Flows // Proceedings of the 32nd International Conference on Machine Learning (ICML). 2015. P. 1530–1538.
10. *Скрипкина Н.В.* Импутация данных муниципальной статистики / Н.В. Скрипкина // Вестник Новосибирского гос. ун-та экономики и управления. 2020. Т. 20. № 4. С. 62–74.
11. *Pedregosa F. et al*. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
12. *Kingma D.P., Welling M.* Auto-Encoding Variational Bayes // Proceedings of the 2nd ICLR. 2014.
13. *Скворцов А.С.* Методы восстановления пропусков в данных: обзор подходов / А.С. Скворцов // Математические методы анализа данных. 2021. Т. 16. № 3. С. 35–60.

14. Импутация неслучайных пропусков в данных опросов с помощью множественной импутации: выпускная квалификационная работа / Нац. исслед. ун-т «Высшая школа экономики». М.: НИУ ВШЭ, 2020.
15. *Rumelhart D.E., Hinton G.E., Williams R.J.* Learning representations by back-propagating errors // Nature. 1986. Vol. 323. P. 533–536.
16. *Ba J.L., Kiros J.R., Hinton G.E.* Layer Normalization // arXiv preprint. 2016. arXiv:1607.06450.
17. *Clevert D.-A., Unterthiner T., Hochreiter S.* Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs) // ICLR. 2016.
18. *Higgins I., Matthey L., Pal A. et al*. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework // ICLR. 2017.
19. *Селянинов Г.Т.* О сельскохозяйственной оценке климата / Г.Т. Селянинов // Труды по сельскохозяйственной метеорологии. 1928. Вып. 20. С. 165–177.
20. *Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D*. Mixup: Beyond Empirical Risk Minimization // Proceedings of the 6th ICLR. 2018.
21. *Efron B.* Bootstrap methods: another look at the jackknife // The Annals of Statistics. 1979. Vol. 7. № 1. P. 1–26.
22. *Efron B., Tibshirani R.J.* An Introduction to the Bootstrap. New York: Chapman & Hall/CRC, 1994. 456 p.
23. *Yoon J., Jordon J., Schaar M.* GAIN: Missing Data Imputation using Generative Adversarial Nets // Proceedings of the 35th ICML. 2018. Vol. 80. P. 5689–5698.
24. Вариационный автоэнкодер для импутации панели агроэкономических данных Кыргызской Республики: результаты обучения, self-masking валидация и диагностика по фактическому прогону пайплайна // Настоящий цикл исследований, статья в печати // Вестник КРСУ. 2026. Т. 26. № 8.
25. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. 2014. Vol. 15. P. 1929–1958.
26. *Gal Y., Ghahramani Z.* Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning // Proceedings of the 33rd ICML. 2016. P. 1050–1059.
27. *Stekhoven D.J., Bühlmann P.* MissForest –non-parametric missing value imputation for mixed-type data // Bioinformatics. 2012. Vol. 28. № 1. P. 112–118.
28. *Скворцова К.А.* Импутация пропусков в официальной статистике: сравнительный анализ методов / К.А. Скворцова // Российский статистический журнал. 2021. Т. 28. № 2. С. 15–33.
29. *Ляликова В.* Восстановление (импутация) данных с помощью Python / В. Ляликова // Habr. 2022. 7 авг. URL: https://habr.com/ru/companies/otus/articles/681410/ (дата обращения: 13.07.2026).
30. Импутация данных: 6 способов восстановить данные в датасете // Tproger. 2020. 23 сен. URL: https://tproger.ru/translations/data-imputation (дата обращения: 13.07.2026).
31. Всеобъемлющее руководство по методам импутации данных в бизнес-аналитике // DataCalculus. 2023. URL: https://datacalculus.com (дата обращения: 13.07.2026).
32. *Скворцов А.С.* Методы восстановления пропусков в данных: обзор подходов // Математические методы анализа данных. 2021. Т. 16. № 3. С. 35–60.
33. Импутация неслучайных пропусков в данных опросов с помощью множественной импутации: выпускная квалификационная работа / Нац. исслед. ун-т «Высшая школа экономики». М.: НИУ ВШЭ, 2020.
34. IBM Corp. Оценка статистики и импутация пропущенных значений: руководство по IBM SPSS Statistics. IBM, 2022. URL: https://www.ibm.com/docs/ru/spss-statistics (дата обращения: 13.07.2026).