НАЦИОНАЛЬНЫЙ СТАТИСТИЧЕСКИЙ КОМИТЕТ
РЕСПУБЛИКИ БЕЛАРУСЬ

СТАТИСТИКА 2030 · ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

От традиционной переписи к data-driven статистике

Белорусский кейс: цифровая трансформация государственной статистики

Проволоцкий Вячеслав Евгеньевич
начальник отдела сопровождения
и разработки информационных систем
РУП «Центр информационных технологий
Национального статистического комитета
Республики Беларусь»

ПАРАДОКС: ИИ ПОВСЮДУ – КРОМЕ ГОССТАТИСТИКИ

ИИ ПРИМЕНЯЕТСЯ ПОВСЕМЕСТНО

- **Промышленность**
- **Медицина**
- **Финансы**
- **Транспорт**
- **Розничная торговля**
- **Маркетинг и реклама**

ГОССТАТИСТИКА ОГРАНИЧИЛАСЬ ЧАТ-БОТАМИ

Виртуальные помощники для респондентов

Поверхностный слой автоматизации

Ядро статпроизводства – сбор, очистка, валидация, интеграция данных остаётся в ручном режиме

Перепись – возможность отработать ИИ не на пилоте, а в реальном крупномасштабном проекте

ЧТО ТАКОЕ DATA-DRIVEN СТАТИСТИКА

СТРАТЕГИЧЕСКАЯ ЦЕЛЬ

Переход от разовых анкетных наблюдений к устойчивой системе государственной статистики, в которой решения опираются на интеграцию и анализ данных из различных источников.

ТРИ ПРИЗНАКА DATA-DRIVEN МОДЕЛИ

01

Единая экосистема источников

Переписи, регистры, административные и большие данные – взаимосвязанные части одной системы

02

Решения на основе данных

Методология, обработка, обновление показателей опираются на системный анализ, а не на экспертную интуицию

03

Правильные данные в момент

Инструменты обеспечивают формирование показателей, оценку качества и своевременный пересмотр методологий

БОЛЕВЫЕ ТОЧКИ ТРАДИЦИОННОЙ ПЕРЕПИСИ

01 **Ошибки в свободных полях**

Адреса, ФИО, названия стран, варианты «другое» – живой язык, не всегда соответствующий классификаторам.

02 **Пропуски и неполные ответы**

Чувствительные вопросы (источники средств, рождаемость), сложные формулировки, технические сбои.

03 **Дубли и противоречия**

Учет по нескольким адресам, расхождения с административными регистрами.

04 **Ручная сверка и кодирование**

Нормализация адресов, перевод стран в коды ISO, разбор вариантов «другое» – существенная часть трудозатрат.

На ручную сверку и кодирование приходится существенная часть трудозатрат этапа обработки

ПОЧЕМУ ПЕРЕПИСЬ – УДОБНАЯ «ПЕСОЧНИЦА» ДЛЯ ИННОВАЦИЙ

01

Разовый проект

Чёткие сроки начала и завершения позволяют оценить эффект «до/после»

02

Измеримые KPI

Доля автоматического кодирования, время до публикации, трудозатраты

03

Политическая поддержка

Приоритетное финансирование и ресурсы для пилотирования технологий

04

Разнообразие данных

Анкеты, регистры, картография, сигнальные данные сотовых операторов

Решения, отработанные на переписи, тиражируются на регулярные статистические наблюдения

АРХИТЕКТУРА ИННОВАЦИЙ: ЧЕТЫРЕ НАПРАВЛЕНИЯ

I

Очистка данных средствами ИИ

Нормализация адресов, ФИО и свободных текстовых полей; автоматическое кодирование вариантов «другое»

II

Обнаружение выбросов и аномалий

Isolation Forest, автоэнкодеры и статистика – подсветка подозрительных записей без автоматического отбраковывания

III

Интеграция с регистром населения

Предзаполнение анкет, импутация пропусков, кросс-валидация с ведомственными источниками

IV

Большие данные сотовых операторов

Сырые обезличенные сигнальные данные: маятниковая миграция, фактическое проживание, плотность населения

Четыре направления работают как единая архитектура – от первичных данных до итоговых показателей

ЛИЦА БЕЗ ЛИЧНОГО ИДЕНТИФИКАТОРА

ПОДАВЛЯЮЩЕЕ БОЛЬШИНСТВО

Имеет личный идентификатор

Связывание с регистрами «жестко» по ключу

ОСТАЛЬНАЯ ЧАСТЬ

Лица без личного идентификатора

Новорожденные, отдельные категории иностранцев и лиц без гражданства

РЕШЕНИЕ – нечеткое сопоставление записей

01

Признаки сопоставления

ФИО, дата рождения, пол, адрес, состав домохозяйства — с учётом опечаток, разных написаний и пропусков

02

Вероятности совпадения

Для каждой пары записей модель рассчитывает вероятность, что они относятся к одному лицу

03

Решение

Высокая вероятность — автообъединение; пограничные — оператору; явные различия — разные лица

Корректный учёт всех лиц без пропусков и двойного счёта

ОГРАНИЧЕНИЯ, РИСКИ И УСЛОВИЯ УСПЕХА

01

Качество исходных данных

Модель не лучше обучающей выборки – нужна системная работа с разметкой

02

Прозрачность и объяснимость

Госстатистика не может позволить «чёрный ящик» – нужно объяснимое обоснование решений

03

Конфиденциальность

Сырые сигнальные данные и регистры требуют обезличивания и юридической проработки

04

Подготовка кадров

Специалисты на стыке анализа данных и предметной области статистики

Эти ограничения – не препятствия, а условия успеха, которые планомерно отрабатываются

ОТ АНКЕТНОЙ МОДЕЛИ —

К СТАТИСТИКЕ, ОСНОВАННОЙ НА ДАННЫХ

01

Полигон для технологий

Перепись — место, где отрабатываются методы и решения

02

Тиражирование на регулярную статистику

Отлаженные решения переходят в постоянные наблюдения

03

Точное и оперативное реагирование

На изменения в социально-экономической среде

СПАСИБО ЗА ВНИМАНИЕ!

НАЦИОНАЛЬНЫЙ СТАТИСТИЧЕСКИЙ КОМИТЕТ
РЕСПУБЛИКИ БЕЛАРУСЬ

https://www.belstat.gov.by/