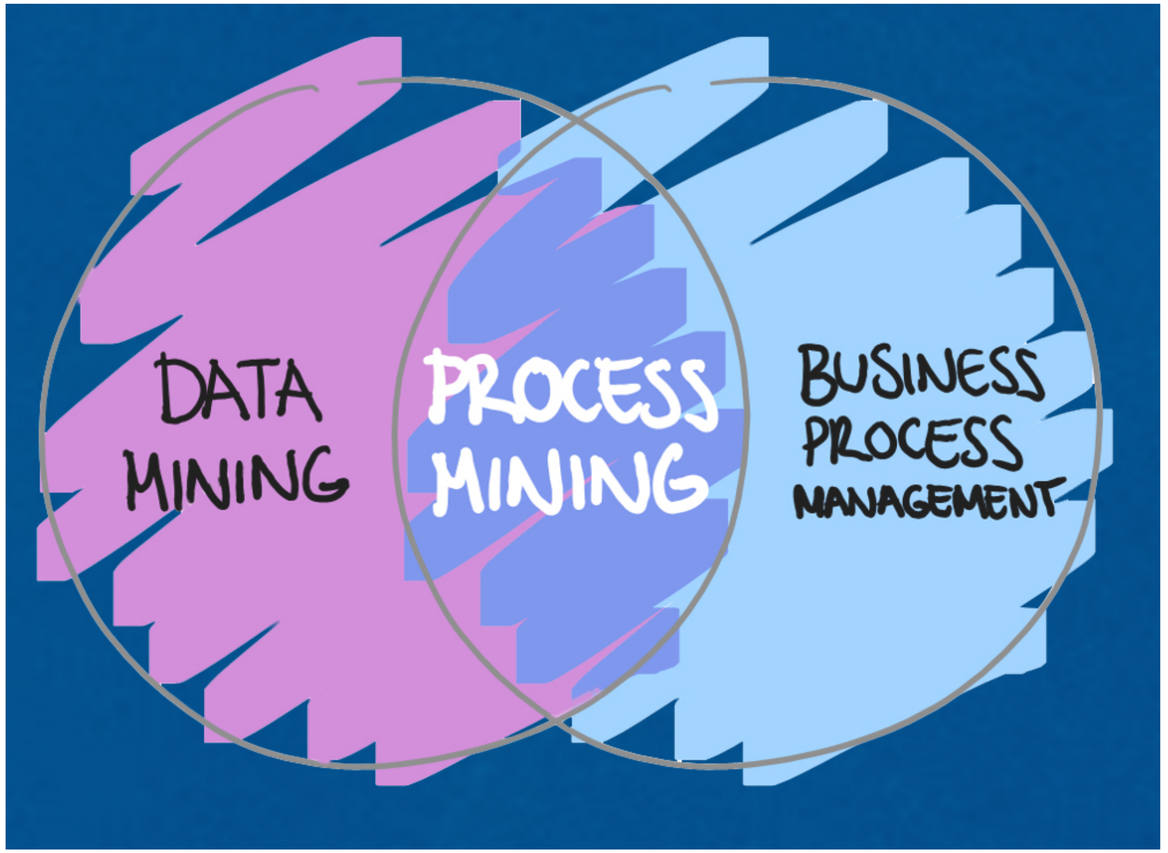
Обнаружение концептуального дрифта в Process mining

работу выполнил:
Посевин М.Э.

Process mining - что это?

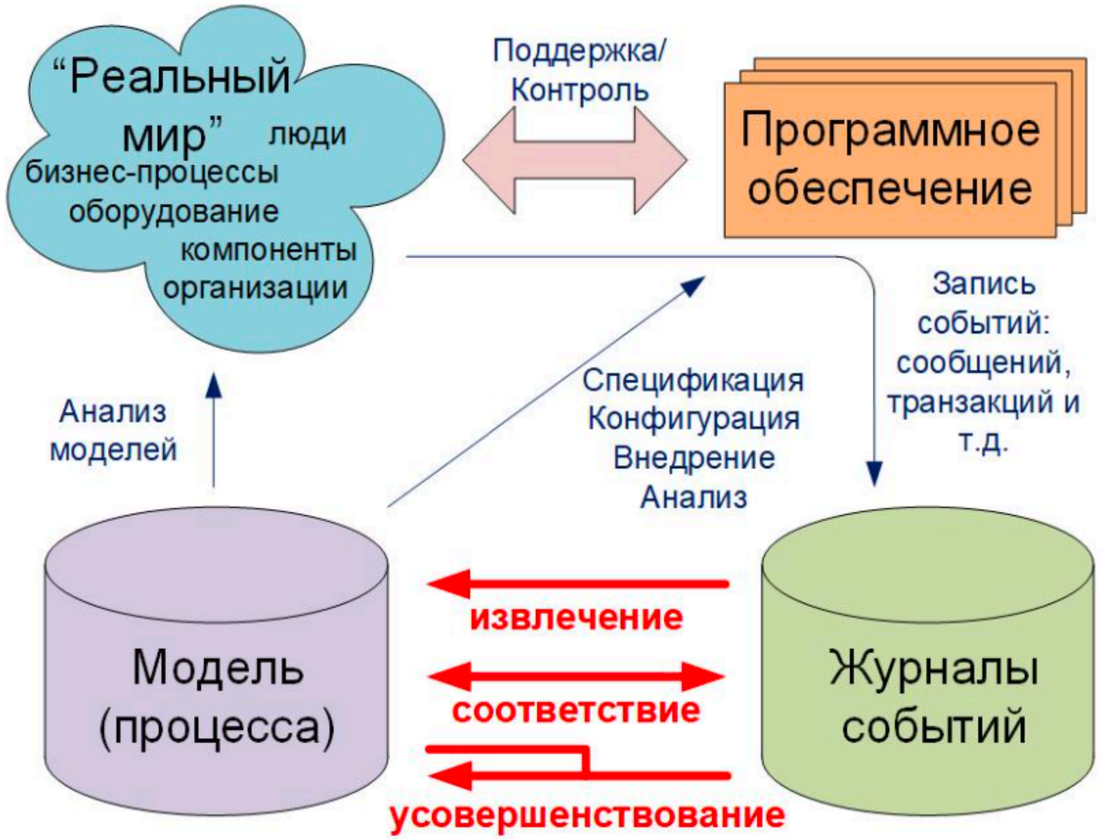
Process mining – сравнительно молодая область исследований, которая находится между направлениями исследований вычислительного интеллекта и глубинным анализом данных, с одной стороны, и моделированием и анализом процессов – с другой.



Process mining - что это?

Process mining - практики, методы и инструменты извлечения, мониторинга и улучшения существующих бизнес-процессов.

В основе process mining лежит извлечение новых знаний из журналов событий, которые массово доступны в современных информационных системах.



Журнал событий

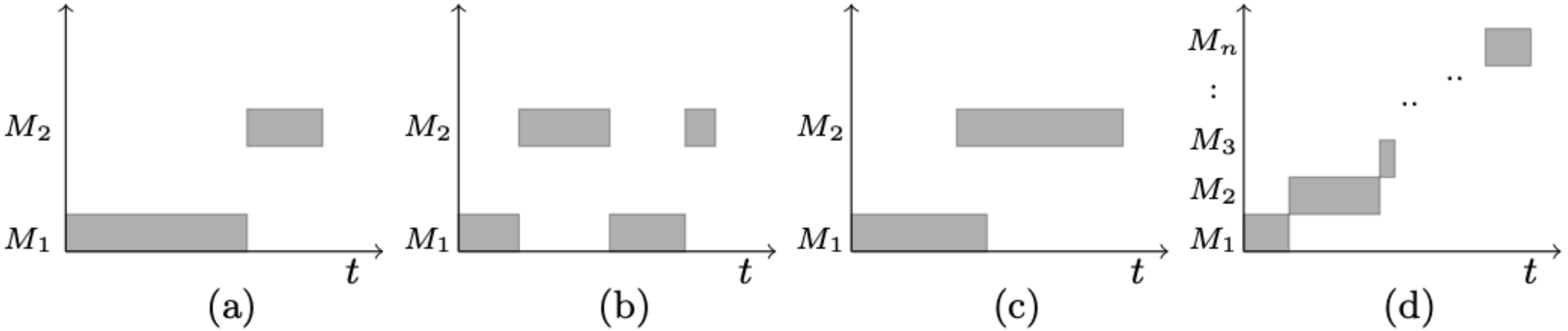
Часто отправной точкой для интеллектуального анализа процессов являются данные из **журналов событий**. Их можно рассматривать как совокупности случаев, а отдельные случаи - как последовательности ссылающихся на них событий.
Выделим основные атрибуты событий:

patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014@10.30	Dr. Jones	45	70.00
5541	blood test	23-1-2014@10.18	Dr. Scott	61	40.00
5833	blood test	23-1-2014@10.27	Dr. Scott	24	40.00
5781	blood test	23-1-2014@10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014@11.10	Dr. Fox	45	1200.00
5833	surgery	23-1-2014@12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014@12.41	Carol Hope	45	0.00
5541	radiation therapy	23-1-2014@13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014@13.08	Dr. Jones	61	140.00
...	...	...	...	...	...

case id | activity name | timestamp | resource | other data

Концептуальный дрифт

Концептуальный дрифт описывает ситуацию, при которой процесс видоизменяется по ходу анализа. Процессы могут видоизменяться в силу периодических/сезонных колебаний или по причине изменившихся условий.



Извлечение признаков

Журналы событий характеризуются взаимоотношениями между событиями. Зависимости между событиями могут быть выражены с помощью *следует* (или *предшествует*) отношений.
Для любой пары событий можно определить, *всегда, никогда* или *иногда* одно из событий *следует* за другим.

$$\mathscr{L} = \{acaebfh, ahijebd, aeghijk\}$$
$$\mathscr{A} = \{a, b, c, d, e, f, g, h, i, j, k\}$$

В $\mathscr{L}$ сохраняются следующие отношения: e *всегда следует* за a, e *никогда не следует* за b, и b *иногда следует* за a

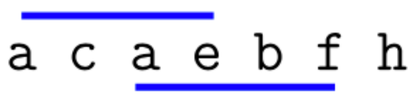
Извлечение признаков

Window count (WC) - функция $f_{WC}^{l,t} : \mathscr{A} \times \mathscr{A} \to \mathbb{N}_0$,

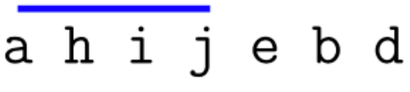
$f_{WC}^{l,t}(a,b) = |\mathscr{F}^{l,t}(a,b)|$,

где $\mathscr{F}^{l,t}(a,b) = [s \in S^{l,t}(a) \,|\, \exists_{1<k\leq|s|} s(k) = b]$,

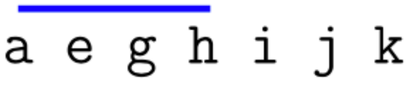
$S^{l,t}(a) = [t(i, i+l-1) \,|\, t \in \mathscr{L}, t(i) = a]$



$f_{\mathrm{WC}}^{4,\mathbf{t}}(\mathtt{a},\mathtt{b}) = 1$



$f_{\mathrm{WC}}^{4,\mathbf{t}}(\mathtt{a},\mathtt{b}) = 0$



$f_{\mathrm{WC}}^{4,\mathbf{t}}(\mathtt{a},\mathtt{b}) = 0$

Извлечение признаков

J-measure - функция $f_J^{l,t} : \mathscr{A} \times \mathscr{A} \to \mathbb{R}^+$,

$$f_J^{l,t}(a,b) = p^t(a)CE^{l,t}(a,b),$$

где $$CE^{l,t}(a,b) = p^{l,t}(a,b)log_2(\frac{p^{l,t}(a,b)}{p^t(b)}) + (1 - p^{l,t}(a,b))log_2(\frac{1 - p^{l,t}(a,b)}{1 - p^t(b)})$$

$$p^{l,t}(a,b) = |\mathscr{F}^{l,t}(a,b)|/|S^{l,t}(a)|$$

Пример: Пусть $t = acaebfh$, $l = 4$

$$p^{4,t}(a,b) = \frac{1}{2}, p^t(a) = \frac{2}{7}, p^t(b) = \frac{1}{7} \Rightarrow f_J^{4,t}(a,b) = 0.147$$

Проверка статистических гипотез

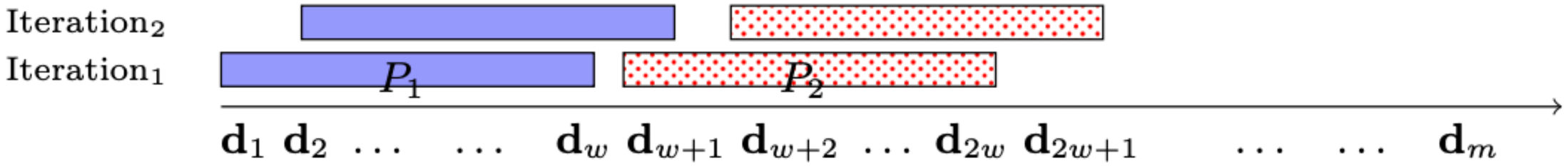
Можно рассматривать журнал событий $\mathscr{L}$ как временную последовательность экземпляров процесса. Разбив журнал на k поджурналов, мы можем преобразовать его в последовательность $\mathscr{D}$, где $\forall d_i \in \mathscr{D}$ соответствует значению выбранного признака для i-го поджурнала.

Предполагается, что должно быть характерное различие в значениях признаков поджурналов до и после точек изменения процесса, причем это различие должно быть более выраженным на границах.

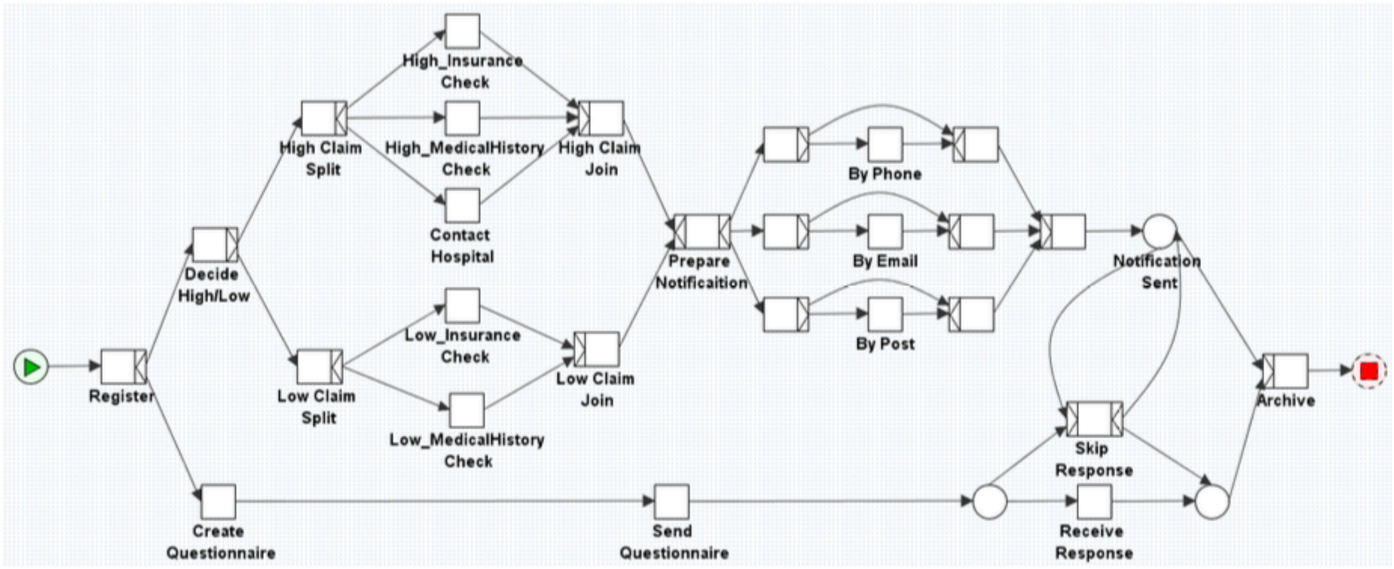
Проверка статистических гипотез

Скользящее окно размера w используется для генерации выборок P_1 и P_2. На каждой итерации проводится проверка статистической гипотезы о принадлежности P_1 и P_2 одному закону распределения (вычисляется p-value).

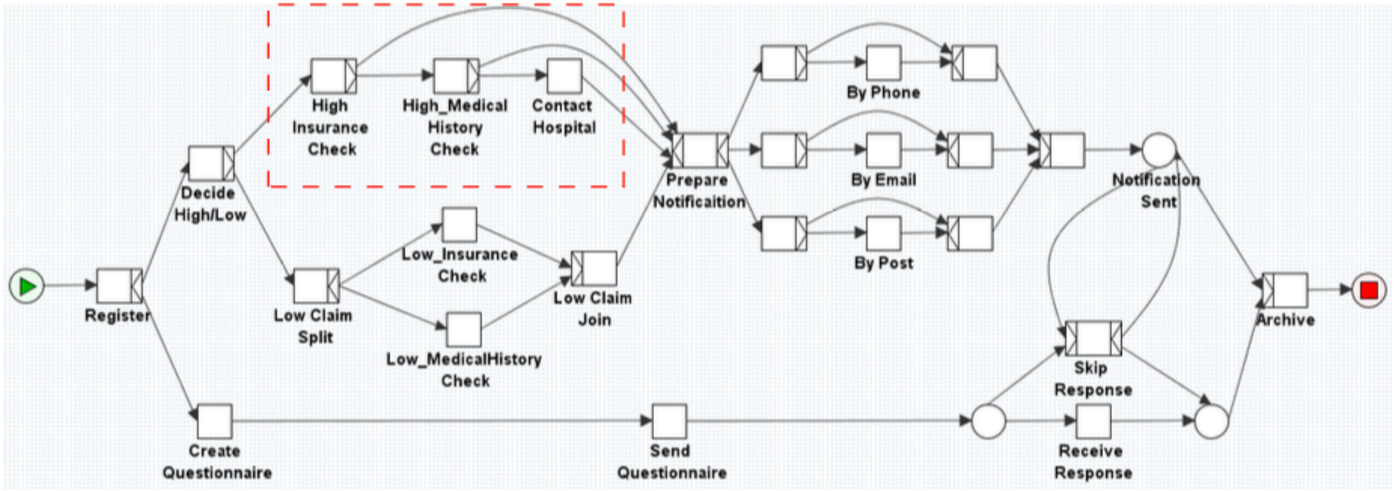
Для этой цели подходит **двухвыборочный критерий однородности Колмогорова-Смирнова**.



Экспериментальные результаты



a. Model, M_1

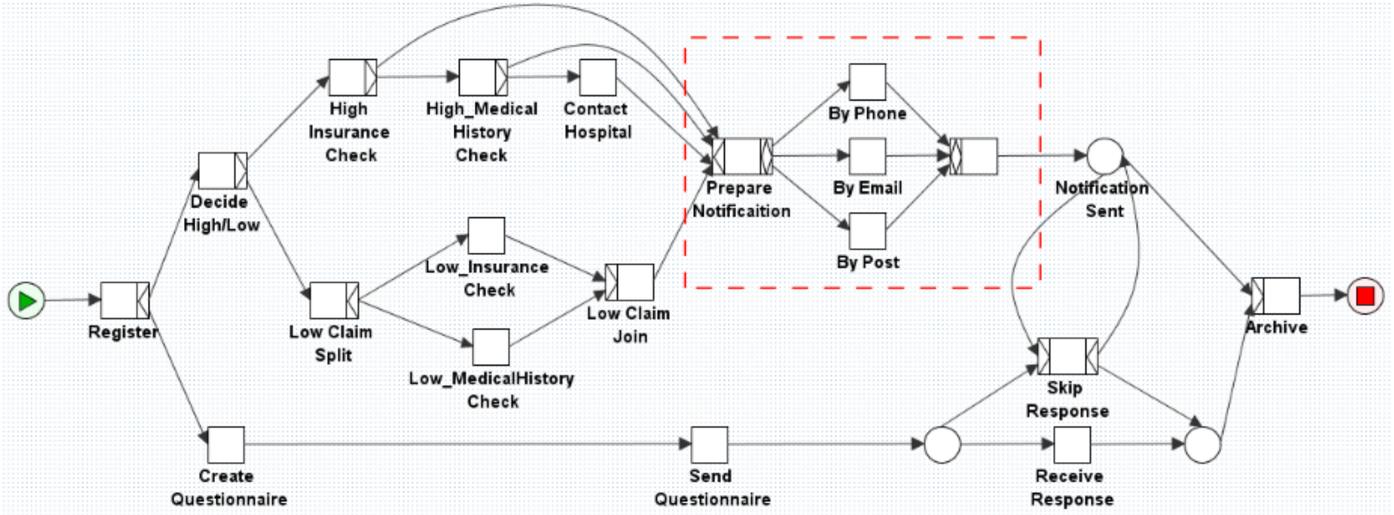


b. Model, M_2

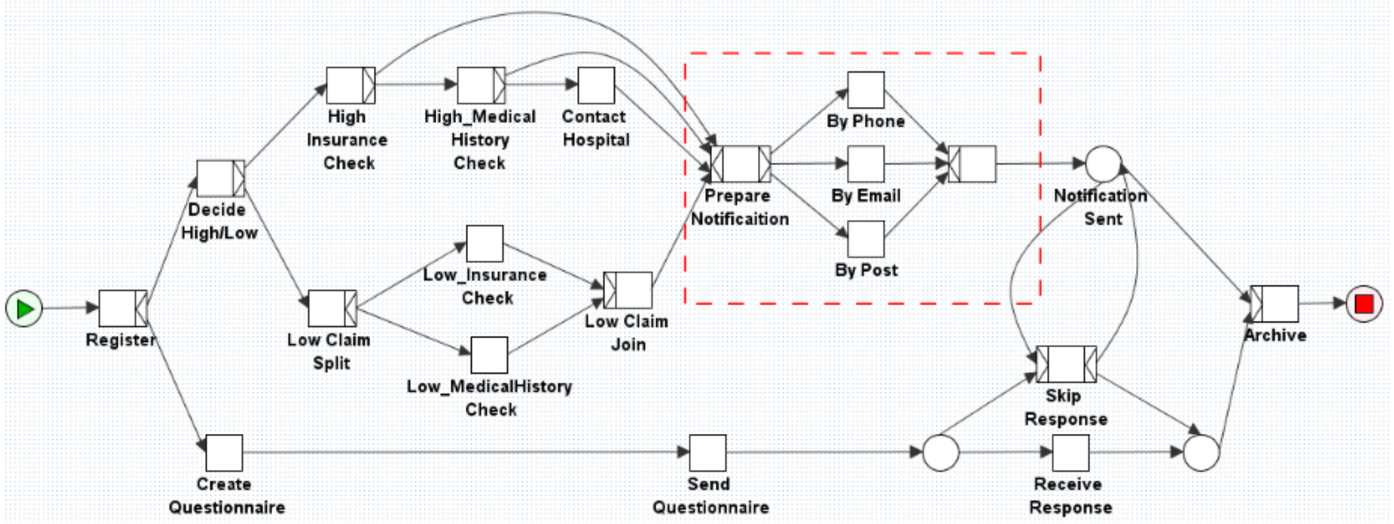
Предложенные идеи будут проиллюстрированы на примере синтетического журнала событий.

Этот журнал содержит информацию о процессе рассмотрения заявлений о медицинском страховании в туристическом агенстве.

Экспериментальные результаты



c. Model, M_3

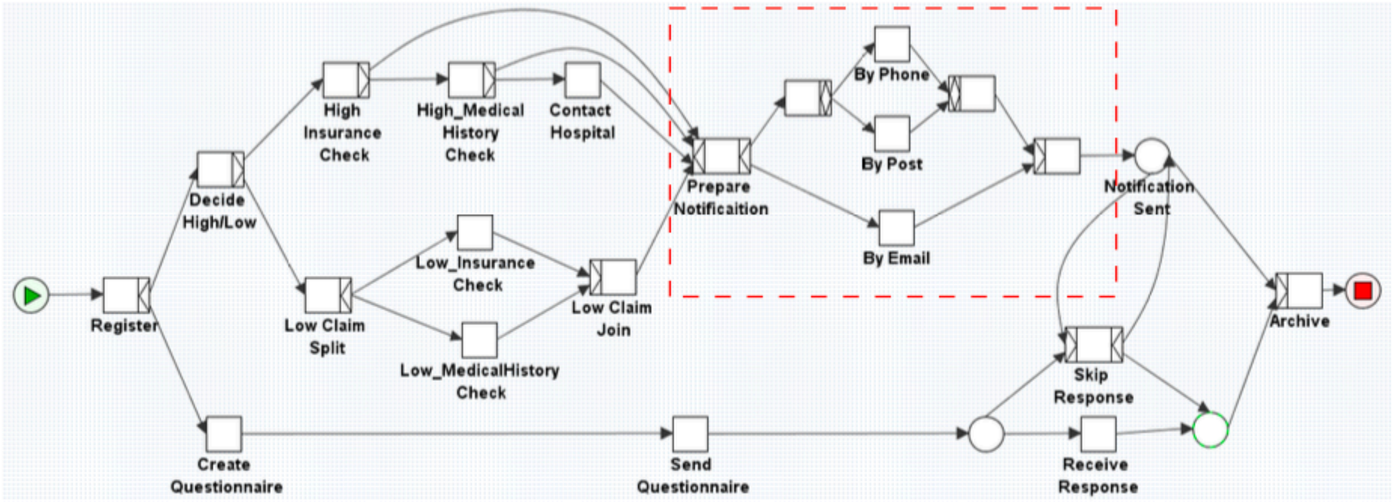


d. Model, M_4

В журнале событий зафиксировано 5 вариантов этого процесса.

Пунктирные прямоугольники указывают на области, в которых были внесены изменения.

Экспериментальные результаты



e. Model, M_5

Обозначим варианты процесса как M1, M2, M3, M4 и M5 (1200 экземпляров процесса для каждой модели)

Журнал событий $\mathscr{L}$ состоит из 6000 экземпляров процесса, содержит 15 видов событий и 58953 событий всего.

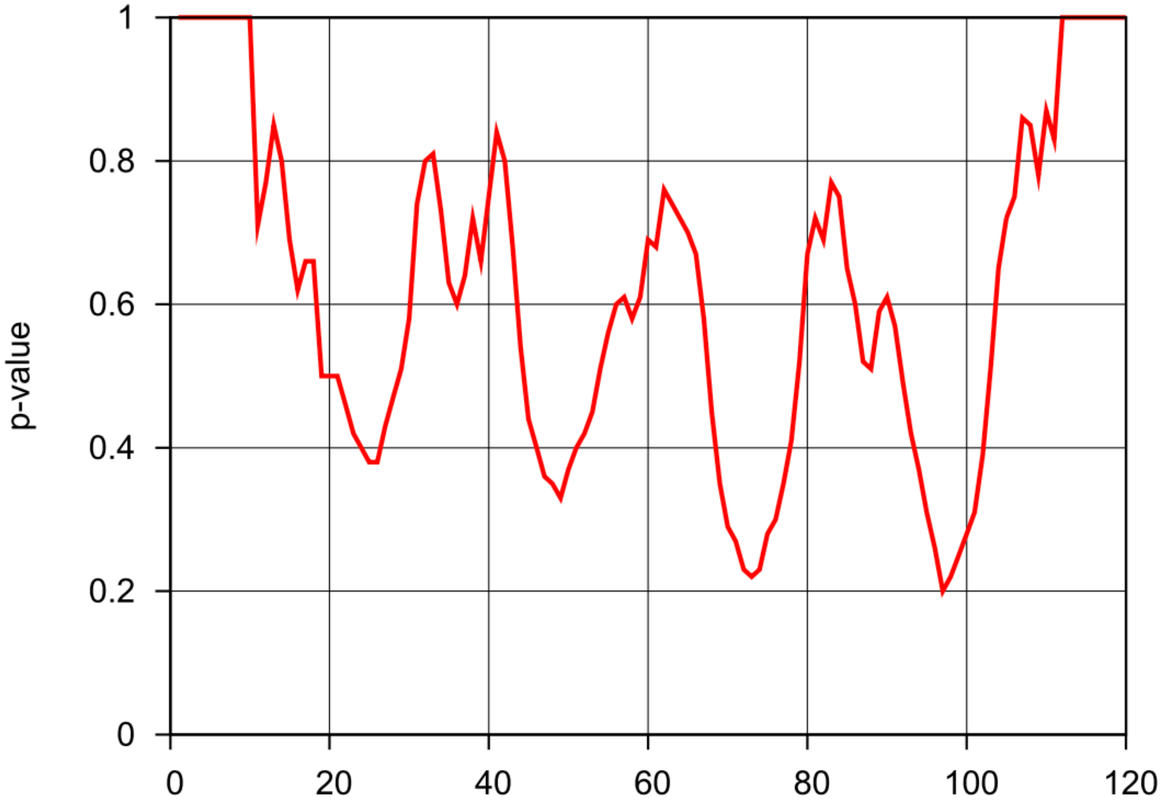
Экспериментальные результаты

1. Разделим журнал на 120 поджурналов по 50 экземпляров процесса.
2. Вычислим J-меру для каждой пары событий на каждой итерации, используя размер окна $w = 10$.
3. Применим критерий однородности Колмогорова-Смирнова к J-мере каждой пары событий с использованием размера окна $w = 10$.
4. Вычислим среднее p-value для каждой итерации.

Экспериментальные результаты

Можно заметить, что на индексах 24, 48, 72 и 96, которые соответствуют фактическим точкам изменения, p-value принимает наименьшее значение.

Что говорит о том, что предложенный метод дает верный результат для данного примера.



Источники

1. R. P. J. C. Bose, W. M. P. van der Aalst, I. Zliobaite, and M. Pechenizkiy. "Dealing with concept drifts in process mining". 2014.

2. Online Course: "Process Mining: Data science in Action". Wil van der Aalst Eindhoven University of Technology, Department Methematics & Computer Science

3. IEEE CIS Task Force on Process Mining. "Process Mining Manifesto". 2011.