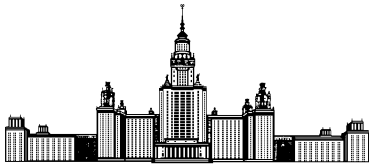
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Фоминская Галина Евгеньевна

Проблема несбалансированности тем в вероятностных тематических моделях

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф-м.н., профессор

Воронцов Константин Вячеславович

Москва, 2019

Содержание

Аннотация

Тематические модели, основанные на матричном разложении и максимизации правдоподобия стремятся выравнивать темы по их мощности, что приводит к дроблению крупных тем и объединению мелких тем в задачах с несбалансированными темами.

В данной работе показано существование проблемы несбалансированности тем. Для решения данной проблемы предложены три подхода: метод итеративной балансировки тем, метод оптимизации гиперпараметров сглаживания и регуляризатор декоррелирования. В экспериментах на полусинтетических данных показано, что первые два метода не решают проблему несбалансированности, а третий метод изменяет балансировку тем, но не восстанавливает исходные темы. Также было показано, что проблема несбалансированности связана с проблемой неоднозначности матричного разложения и выбора начального приближения.

1 Введение

Тематическое моделирование [1], [2] — одно из приложений машинного обучения к анализу текстов. В тематической модели коллекции текстовых документов каждая тема определяется как дискретное распределение на множестве терминов, а каждый документ — как дискретное распределение на множестве тем. Предполагается, что каждый документ — набор терминов, выбранных независимо и случайно из смеси распределений. Задача тематического моделирования состоит в восстановлении компонент смеси по выборке.

Таким образом, построение тематической модели сводится к решению некорректно поставленной задачи неотрицательного матричного разложения. Для решения этой задачи используется максимизация логарифма правдоподобия с помощью EM-алгоритма. Некорректность этой задачи заключается в том, что множество ее решений в общем случае бесконечно и EM-алгоритм может сходиться к локальному оптимуму правдоподобия. Для нахождения конкретного решения и формализации дополнительных требований к модели используется аддитивная регуляризация [3].

Тематической модели, основанной на максимизации правдоподобия, выгодно делать темы одинаковыми по мощности, в то время как реальные пропорции тем определяются историей формирования коллекции. Это приводит к дроблению крупных тем, слиянию мелких тем и утрате интерпретируемости тем. Мы назвали это проблемой несбалансированности тем. В данной работе экспериментально показано существование данной проблемы. Для решения данной проблемы предложены три подхода: метод итеративной балансировки тем, метод оптимизации гиперпараметров сглаживания и регуляризатор декоррелирования. Показано, что регуляризатор декоррелирования восстанавливает баланс тем, но без соответствия с исходными темами. Так же было показано, что проблема несбалансированности связана с проблемой неоднозначности матричного разложения.

1.1 Постановка задачи тематического моделирования

1.1.1 Вероятностный латентный семантический анализ (PLSA)

Пусть даны D — коллекция текстовых документов и W — словарь коллекции, то есть множество слов, встречающихся в этих документах. Для каждого документа d и для каждого слова w известно n_{dw} — сколько раз слово w встретилось в документе d.

Введем также обозначения n — число словопозиций в коллекции, n_d — в документе d.

Будем считать, что выполнены следующие предположения:

- Существует конечное множество T скрытых переменных, называемых темами, и каждое появление слова w в документе d связано с некоторой темой $t \in T$.
- Гипотеза «мешка слов»: порядок расположения слов в документе не имеет значения.
- Гипотеза условной независимости: вероятность принадлежности слова к теме не зависит от документа, в котором встретилось это слово.

$$p(w|d,t) = p(w|t).$$

Решается задача нахождения стохастических матриц распределений слов в темах Φ и тем в документах Θ, таких, что $\varphi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$. В модели PLSA для решения этой задачи используется максимизация правдоподобия:

$$\mathcal{L} = \prod_{d \in D, w \in d} p(w|d)^{n_{dw}} \to \max_{\Phi,\Theta},$$

где $p(w|d) = \sum\limits_{t \in T} p(w|t)p(t|d) = \sum\limits_{t \in T} \varphi_{wt}\theta_{td}$ — вероятностная тематическая модель.

$$L(\Phi,\Theta) = \sum_{d \in D, w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0.$$

Эта задача решается с помощью EM-алгоритма.

Часто для оценки качества тематической модели используется перплексия (perplexity):

$$perp(D;p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right).$$

Чем меньше перплексия, тем лучше модель предсказывает появление слова w в документе d. Мы будем следить за сходимостью перплексии модели с итерациями обучения, чтобы определить момент, когда модель сошлась.

1.1.2 Аддитивная регуляризация тематических моделей (ARTM)

Можно заметить, что исходная постановка задачи некорректна, так как существует бесконечное число решений. Для нахождения конкретного решения и формализации дополнительных требований к модели используется аддитивная регуляризация:

$$L(\Phi,\Theta) = \sum_{d\in D, w\in d} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

где

$$R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta).$$

Коэффициенты регуляризации τ_i обычно подбираются экспериментально по выбранному критерию качества.

1.2 Несбалансированность тем

Мощностью темы будем считать значение $p(t) = \frac{1}{n}\sum\limits_{d\in D} p(t|d)n_d$.

Степенью несбалансированности коллекции назовём отношение максимальной мощности темы в коллекции к минимальной: $k = \frac{\max\limits_t p(t)}{\min\limits_t p(t)}$.

Тематической модели, основанной на максимизации правдоподобия, выгодно делать темы одинаковыми по мощности, в то время как реальные пропорции тем определяются историей формирования коллекции. Это приводит к дроблению крупных тем, слиянию мелких тем и утрате интерпретируемости тем (рис. 1). В коллекции, где 980 документов по биологии, 10 по математике и 10 по социологии, в тематической модели с 3 темами скорее всего все три темы будут по биологии. В модели со 100 темами будет 98 биологических тем, близких друг к другу, и одна тема по

математике и одна по социологии, при этом последние две будут сильно отличаться от остальных.

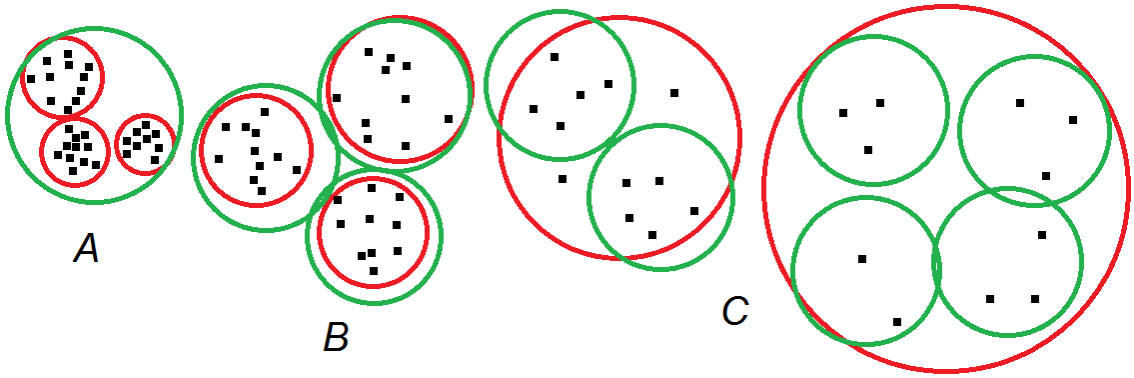


Рис. 1: Тематические модели стремятся выравнивать темы по их мощности (красные кластеры). Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C). Лишь часть тем оказываются семантически однородными и нераздробленными (B).

1.3 Цели и задачи

Цель данной работы в том, чтобы показать существование проблемы несбалансированности тем и найти метод решения проблемы.

1.4 Регуляризатор декоррелирования

Потребуем, чтобы модель находила как можно более различные темы. В [5], [3] для этого было предложено уменьшать ковариации тем:

$$R_2(\Phi, \Theta) = -\frac{1}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws} \to \max_{\Phi, \Theta}.$$

Известно, что этот регуляризатор имеет несколько полезных побочных эффектов: он разреживает столбцы матрицы Φ, способствует выделению слов общей лексики в отдельные «фоновые» темы и улучшает интерпретируемость тем. В данной работе мы проверяем гипотезу, что декоррелирование также восстанавливает «естественный» баланс тем.

1.5 Регуляризаторы сглаживания/разреживания

Потребуем от модели, чтобы столбцы φ_t были близки к заданным распределениям $\beta_t = (\beta_{wt})_{w \in W}$. В работе [4] для этого был предложен регуляризатор сглаживания матрицы Φ, который состоит в минимизации перекрестной энтропии между этими распределениями:

$$\sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} \to \max_{\Phi}.$$

Аналогично был введён регуляризатор сглаживания столбцов матрицы Θ, а также было показано, что при отрицательном коэффициенте регуляризации минимизация перекрестной энтропии переходит в максимизацию, а регуляризаторы становятся разреживающими.

1.6 Оптимизация гиперпараметров сглаживания Θ

В [7] чтобы учесть то, что темы могут входить в коллекцию в разных пропорциях, было предложено вводить асимметричное априорное распределение Дирихле на столбцы матрицы Θ. На языке аддитивной регуляризации метод сводится к оптимизации коэффициентов регуляризаторов сглаживания (при $\alpha_t > 0$) или разреживания (при $\alpha_t < 0$):

$$R(\Theta) = \sum_{d \in D} \sum_{t \in T} \alpha_t \log \theta_{td} \to \max_{\Theta}.$$

Для итерационного обновления коэффициентов α_t в [7] предложена формула:

$$\alpha_t := \alpha \frac{n_t + \frac{\alpha'}{|T|}}{\sum_{t \in T} n_t + \alpha'},$$

где α и α' — новые параметры метода.

2 Итеративное балансирование тем

Рассмотрим задачу максимизации регуляризированного log-правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \ \to \ \max_{\Phi, \Theta} \tag{1}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{2}$$

В [3] показано, что точка локального экстремума этой задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t \mid d, w)$:

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{3}$$

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \tag{4}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}. \tag{5}$$

Значение $n_{tdw} = n_{dw} p_{tdw}$ является оценкой числа вхождений слова w в документ d, связанных с темой t. Для устранения несбалансированности тем предлагается домножить n_{tdw} на величину, обратно пропорциональную n_t.

Допустим, что n_{tdw} увеличилось в k_t раз по всей коллекции: $n'_{tdw} = k_t n_{tdw}$. Тогда вероятности p_{tdw} модифицируются следующим образом:

$$p'_{tdw} = \frac{n'_{tdw}}{\sum_s n'_{sdw}} = \frac{k_t n_{tdw}}{\sum_s k_s n_{sdw}} = k_t p_{tdw} \frac{\sum_s n_{sdw}}{\sum_s k_s n_{sdw}} = \underset{t \in T}{\operatorname{norm}}\big(k_t p_{tdw}\big).$$

Положим $k_t = \frac{1}{n_t}$, где n_t — мощность темы t, вычисленная на предыдущей итерации EM-алгоритма.

Мощности тем n_t будем оценивать через немодифицированные вероятности p_{tdw}, а параметры модели — через модифицированные p'_{tdw}.

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \qquad n_t = \sum_{d \in D}\sum_{w \in d} n_{dw} p_{tdw}.$$

$$p'_{tdw} = \underset{t \in T}{\operatorname{norm}}\left(\frac{\varphi_{wt}\theta_{td}}{n_t}\right); \tag{6}$$

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d \in D} n_{dw} p'_{tdw};$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w \in d} n_{dw} p'_{tdw}.$$

3 Вычислительные эксперименты

Эксперименты были проведены на коллекции postnauka (3446 документов небольшого размера). Предварительно была проведена лемматизация, были удалены стоп-слова. Для построения моделей использовалась библиотека BigARTM [6].

3.1 Генерация коллекций различной степени сбалансированности

Для того, чтобы показать существование проблемы, построим линейку полусинтетических коллекций разной степени несбалансированности.

На коллекции postnauka строится модель с 20 темами с регуляризатором декорреляции столбцов матрицы Φ с коэффициентом $\tau = 10^{10}$. Коллекция состоит из коротких документов, можем считать, что большинство из них монотематичные. С помощью построенной модели выбираем монотематичные документы по следующему принципу: документ d будем считать монотематичным, если $p(t_1|d) > 2p(t_2|d)$, где t_1, t_2 — две наибольшие темы в этом документе.

Из всех монотематичных документов коллекции оставим только те, которые удовлетворяют гипотезе условной независимости. Для этого для каждого монотематичного документа d и его темы t строится эмпирическое распределение статистики Кресси-Рида следующим образом: из распределения φ_t генерируется 1000 псевдо-документов, каждый длиной n_d слов. Для каждого из этих псевдо-документов $\hat{d}$ и темы t вычисляется значение $CR(t, \hat{d})$. По полученному набору значений строится эмпирическое распределение. Для него вычисляется R^{α}_{dt} — $\alpha-$квантиль, $\alpha = 0.95$. Если $CR(t, d) \leq R^{\alpha}_{dt}$, то документ d удовлетворяет гипотезе условной независимости и мы его оставляем.

Из отобранных монотематичных документов будем составлять новые коллекции. Так как теперь документы монотематичные, мы считаем, что для каждой темы $p(t) = \frac{1}{n}\sum_d n_d[d \in t]$. Для получения коллекции с заданным $p^*(t)$ для каждого слова w в документе d из темы t изменим $n_{wd} := n_{wd}\frac{p^*(t)}{p(t)}$. Для желаемого коэффициента несбалансированности k $p^*(t)$ построим следующим образом:

$$p^*(t) = \frac{t^\gamma}{\sum_{t=1}^{T} t^\gamma}, \quad \gamma = \frac{\log k}{\log T}.$$

На рисунке 2 показаны полученные распределения $p(t)$ в синтезированных коллекциях.

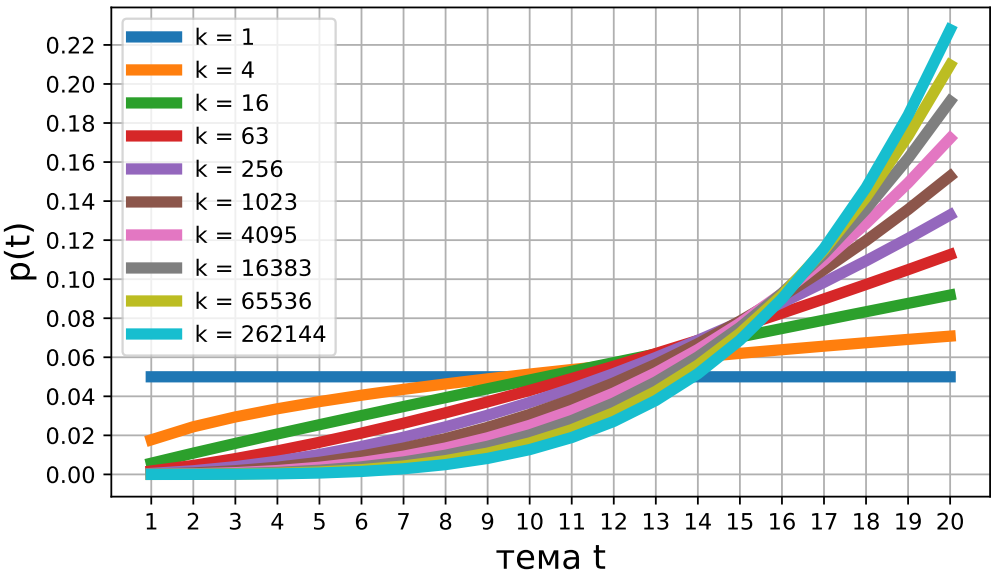


Рис. 2: распределения p(t) для сгенерированных коллекций

3.2 Выявление проблемы несбалансированности

Цель этого эксперимента в том, чтобы показать существование проблемы несбалансированности тем. Для этого на каждой из сгенерированных ранее коллекций построим тематическую модель PLSA без регуляризации и сравним распределение мощностей полученных тем с тем распределением, которое известно по построению этих коллекций.

На рисунке 3 показаны полученные распределения. Видно, что модель PLSA строит примерно равномощные темы независимо от того, насколько неравномерным было распределение мощностей тем по построению коллекции. Для тем в новой модели можно построить соответствие с исходными темами, используя венгерский алгоритм. На рисунке 4 показаны полученные распределения, но новые темы упорядочены в соответствии со старыми.

3.3 Декоррелятор

Цель этого эксперимента в том, чтобы проверить, может ли декоррелятор решить проблему несбалансированности. Декоррелятор включался после 50 итераций на ещё 30 итераций с коэффициентом регуляризации 10^{20}. На рисунке 5 показаны

результаты восстановления распределения мощностей тем для модели ARTM с регуляризатором декоррелирования. На рисунке 6 те же распределения, но новые темы не отсортированы в порядке возрастания мощности, а сопоставлены исходным темам.

Видно, что декоррелятор делает распределение мощностей тем менее равномерным. Но при этом из рисунка 6 видно, что при восстановлении соответствия исходных и новых тем распределение мощностей не сохраняется.

Таким образом, декоррелятор не может полностью решить проблему несбалансированности тем.

3.4 Итеративная перебалансировка тем

На рисунке 7 показаны полученные распределения. На рисунке 8 показаны полученные распределения, но новые темы упорядочены в соответствии со старыми. Видно, что итеративная перебалансировка тем не дает желаемого эффекта.

3.5 Оптимизация гиперпараметров сглаживания Θ

На рисунке 9 показаны полученные распределения. Видно, что рассмотренный метод не решает проблему несбалансированности. На рисунке 10 показаны полученные распределения, но новые темы упорядочены в соответствии со старыми.

3.6 Начальная инициализация Φ

Цель этого эксперимента в том, чтобы разделить проблему несбалансированности и проблему неоднозначности матричного разложения. В модели PLSA матрица Φ была проинициализирована матрицей Φ, использовавшейся при построении коллекций. На рисунке 11 показаны полученные распределения. На рисунке 12 показаны полученные распределения, но новые темы упорядочены в соответствии со старыми. Видно, что распределение $p(t)$ и сами темы хорошо восстановились. Можно сделать вывод, что одна и та же модель может восстановить разную степень несбалансированности тем в зависимости от начального приближения.

3.7 Обсуждение и выводы

В ходе экспериментов было показано существование проблемы несбалансированности, что согласуется с исходным предположением. Было показано, что рассмотренные методы не могут решить эту проблему. Вопрос о том, как строить тематическую модель, чтобы восстанавливались именно исходные темы в исходных пропорциях (и возможно ли это вообще), остался открытым. Было показано, что тематическая модель в зависимости от начального приближения может восстанавливать разные распределения тем по мощностям. Из этого можно сделать вывод, что проблема несбалансированности связана с проблемой неоднозначности матричного разложения.

4 Результаты, выносимые на защиту

- Показано, что методы вероятностного тематического моделирования сталкиваются с проблемой несбалансированности тем
- Предложены два способа балансировки тем, но оба они не приводят к желаемому результату в экспериментах на синтетических данных
- Показано, что получение несбалансированных тем возможно путём выбора начального приближения

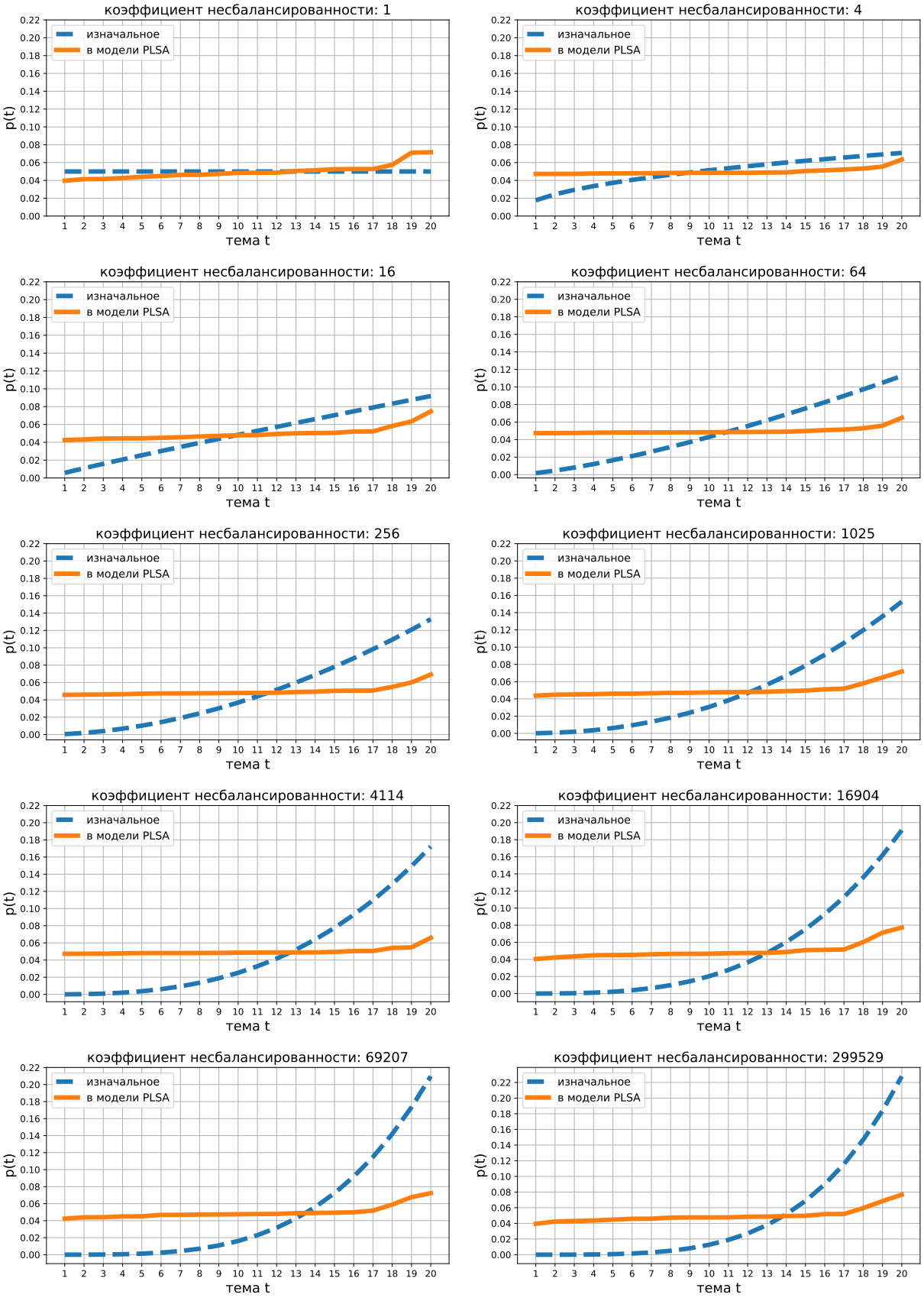


Рис. 3: распределение p(t), полученное PLSA, без сопоставления тем.

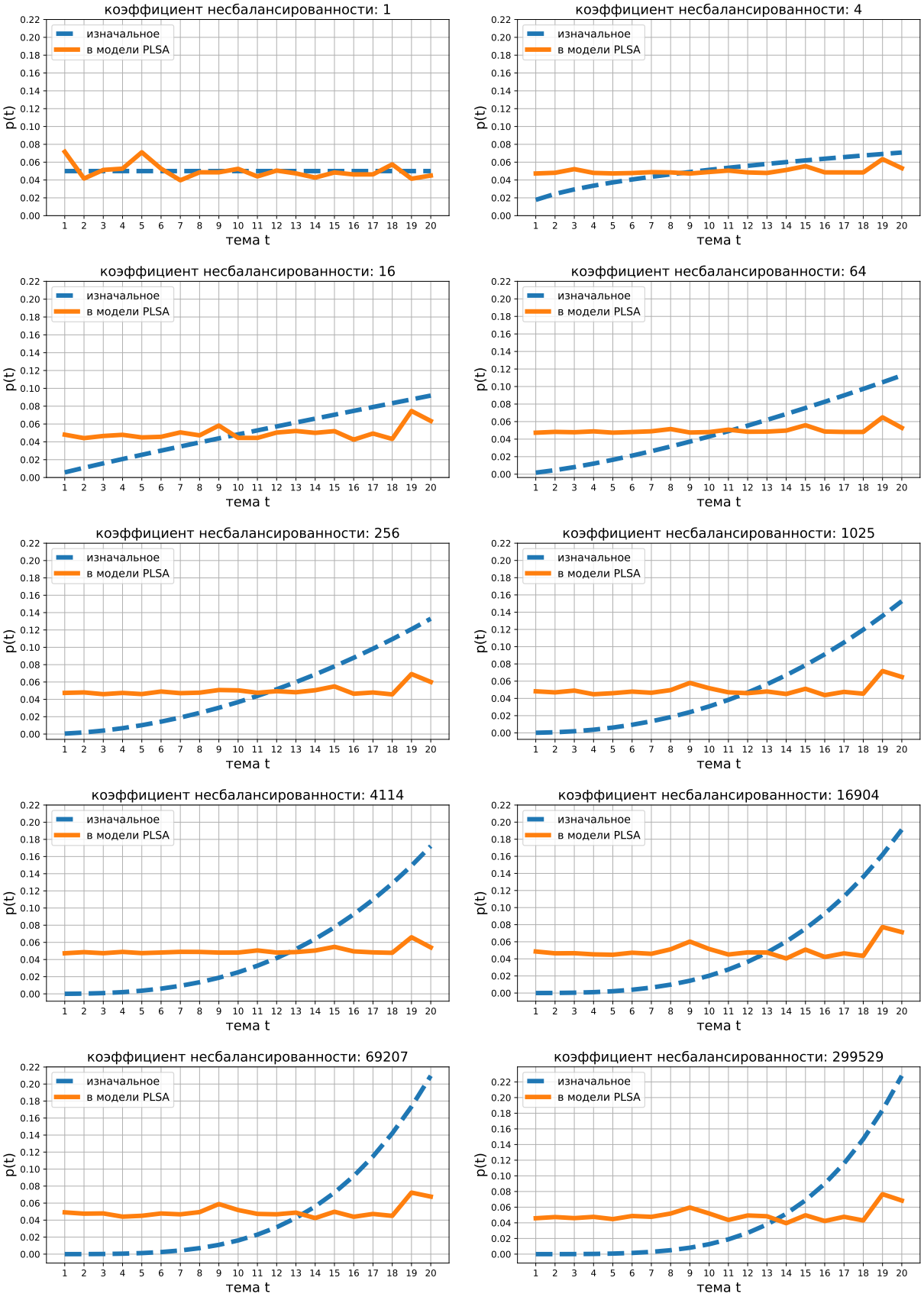


Рис. 4: распределение p(t), полученное PLSA, новые и исходные темы сопоставлены венгерским алгоритмом.

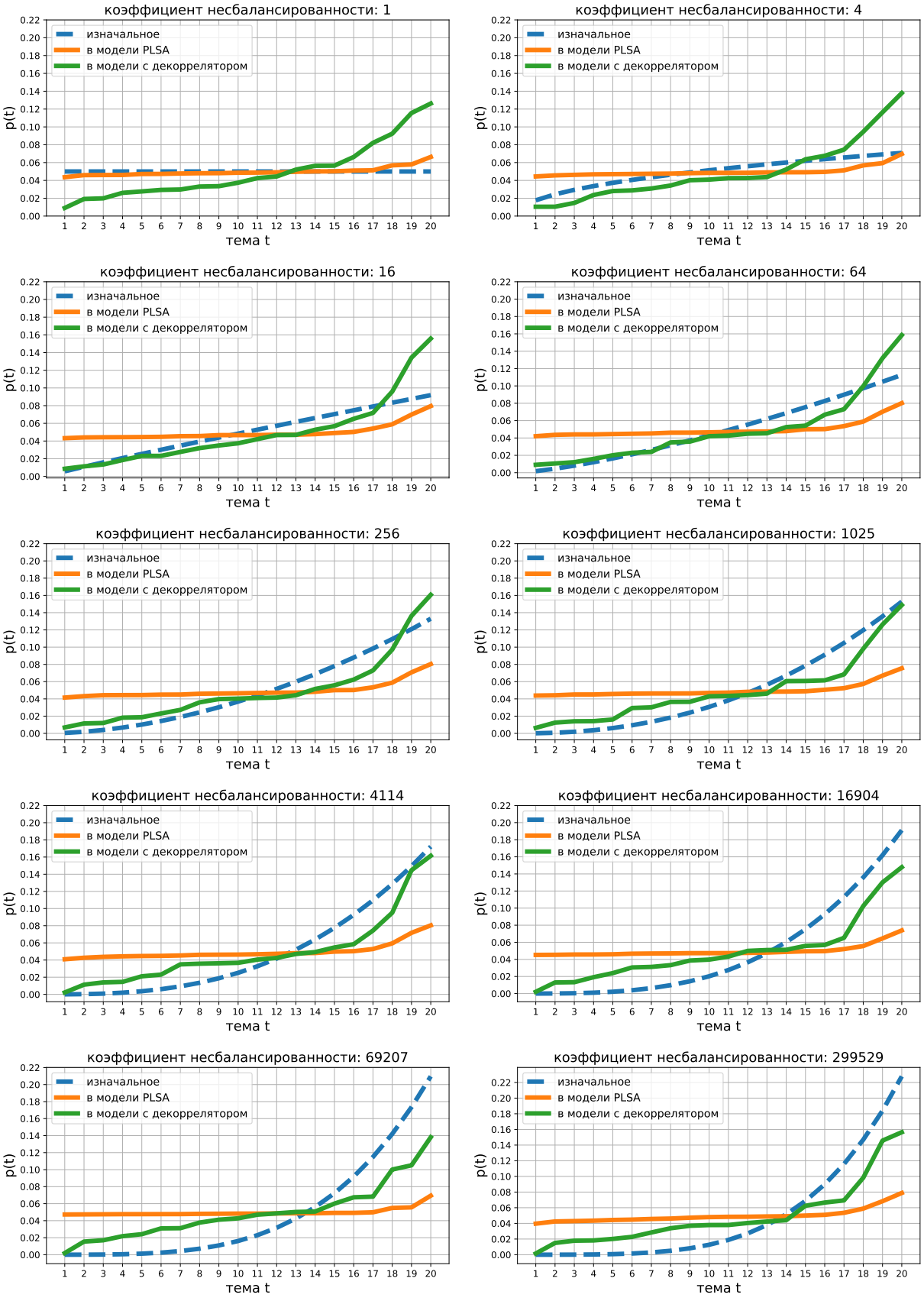


Рис. 5: распределение p(t), полученное моделью с декоррелятором, без сопоставления тем.

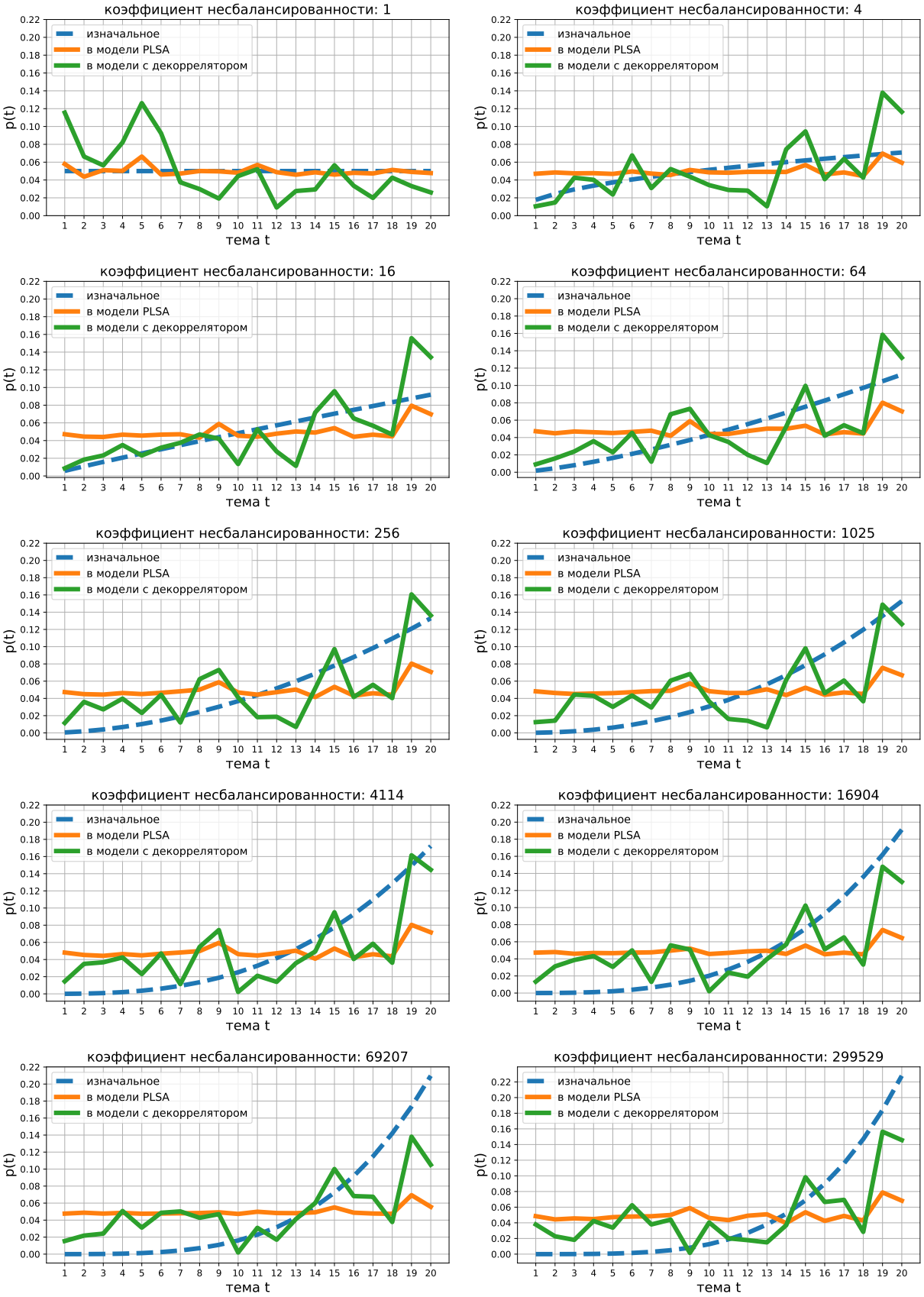


Рис. 6: распределение p(t), полученное моделью с декоррелятором, новые и исходные темы сопоставлены венгерским алгоритмом.

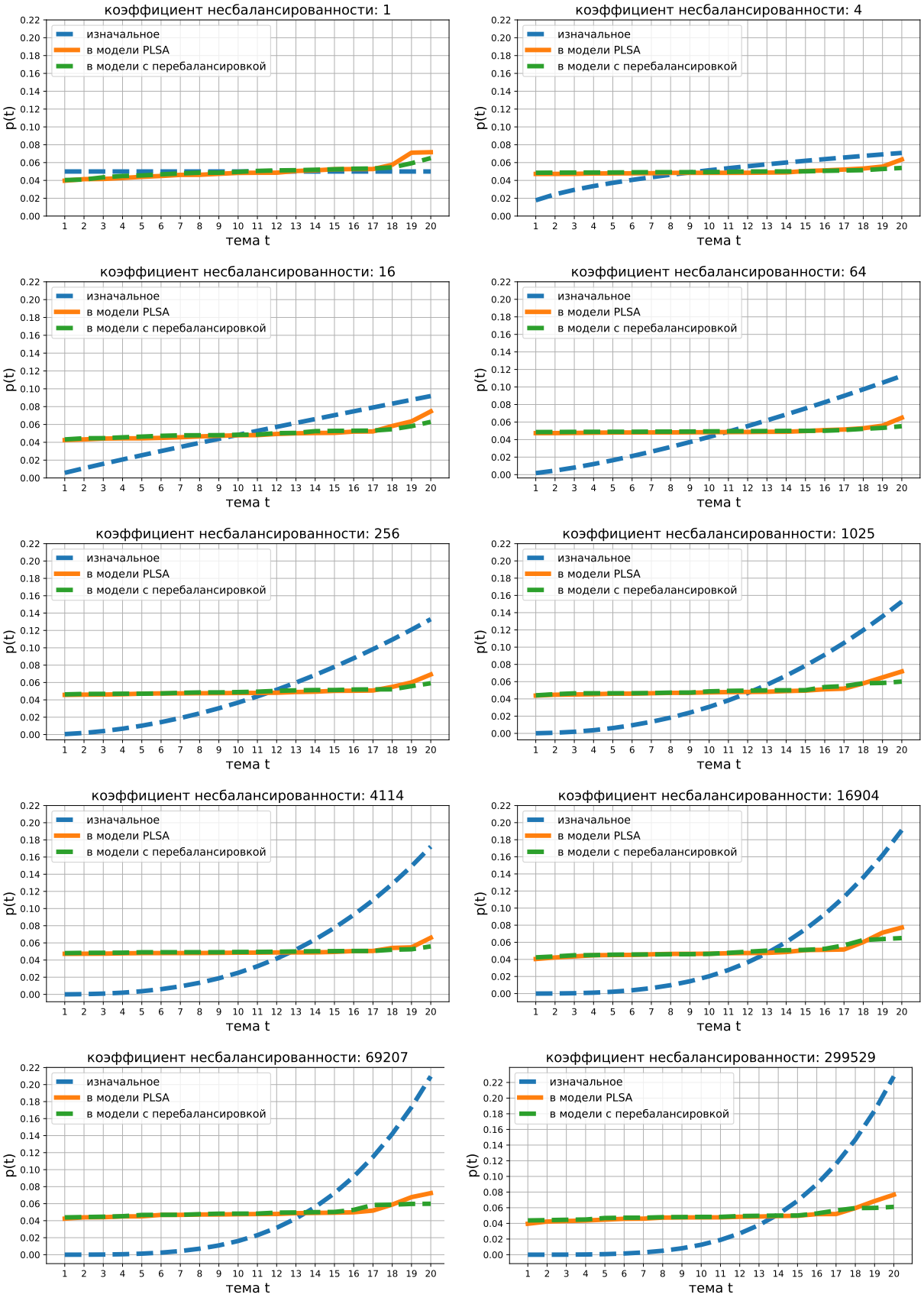


Рис. 7: распределение p(t), полученное моделью с балансировкой, без сопоставления тем.

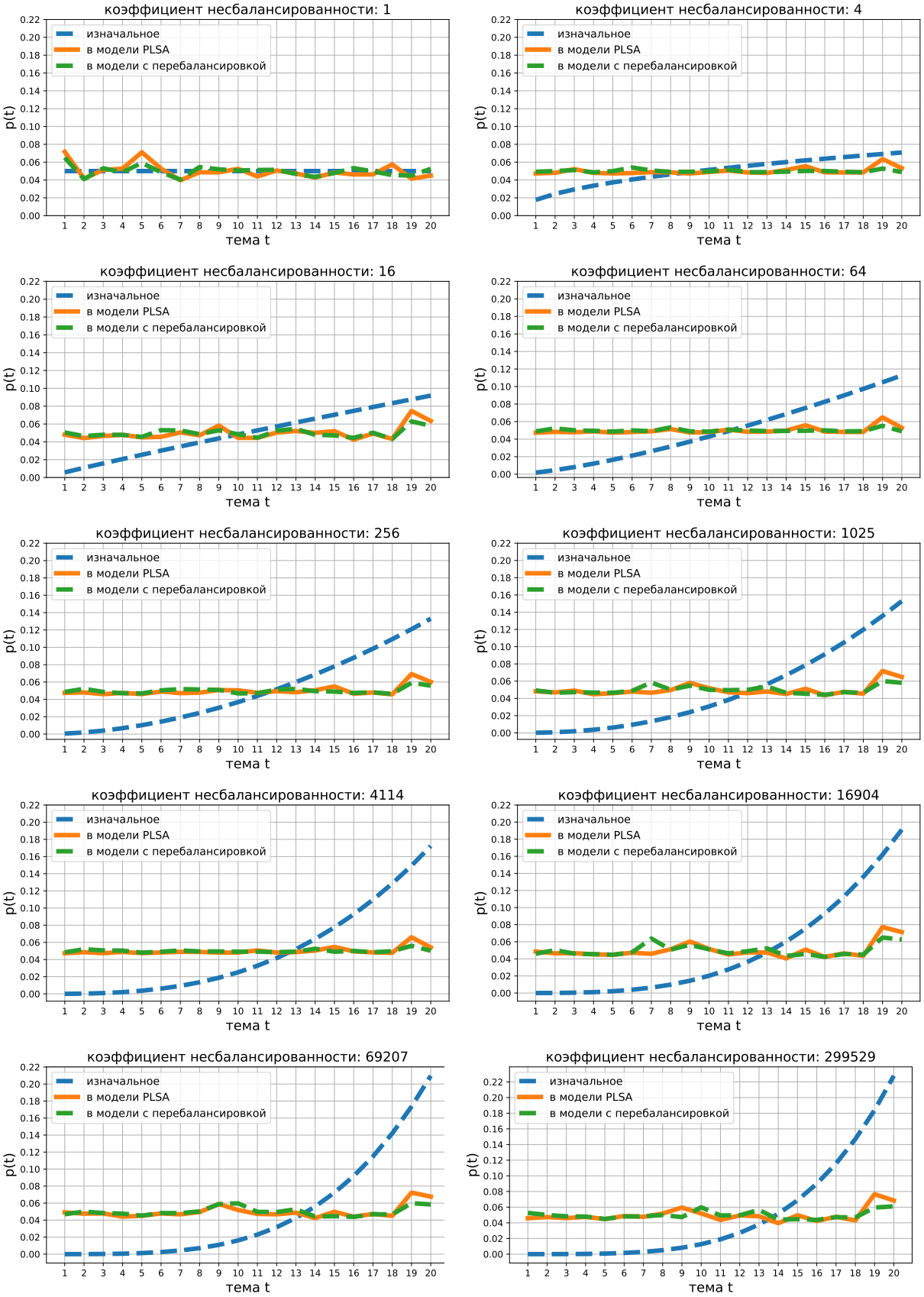


Рис. 8: распределение p(t), полученное моделью с балансировкой, новые и исходные темы сопоставлены венгерским алгоритмом.

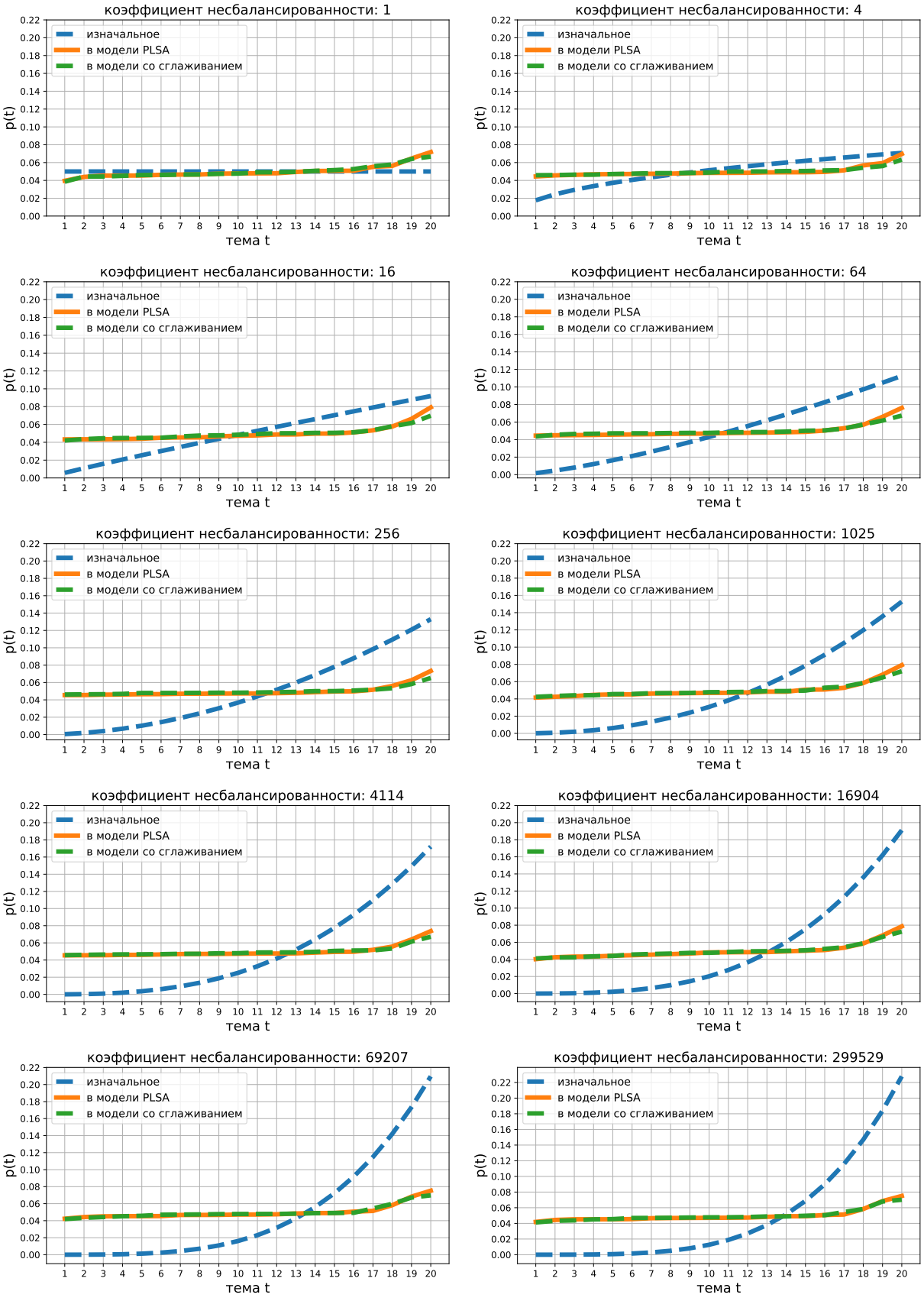


Рис. 9: распределение p(t), полученное моделью со сглаживанием, без сопоставления тем.

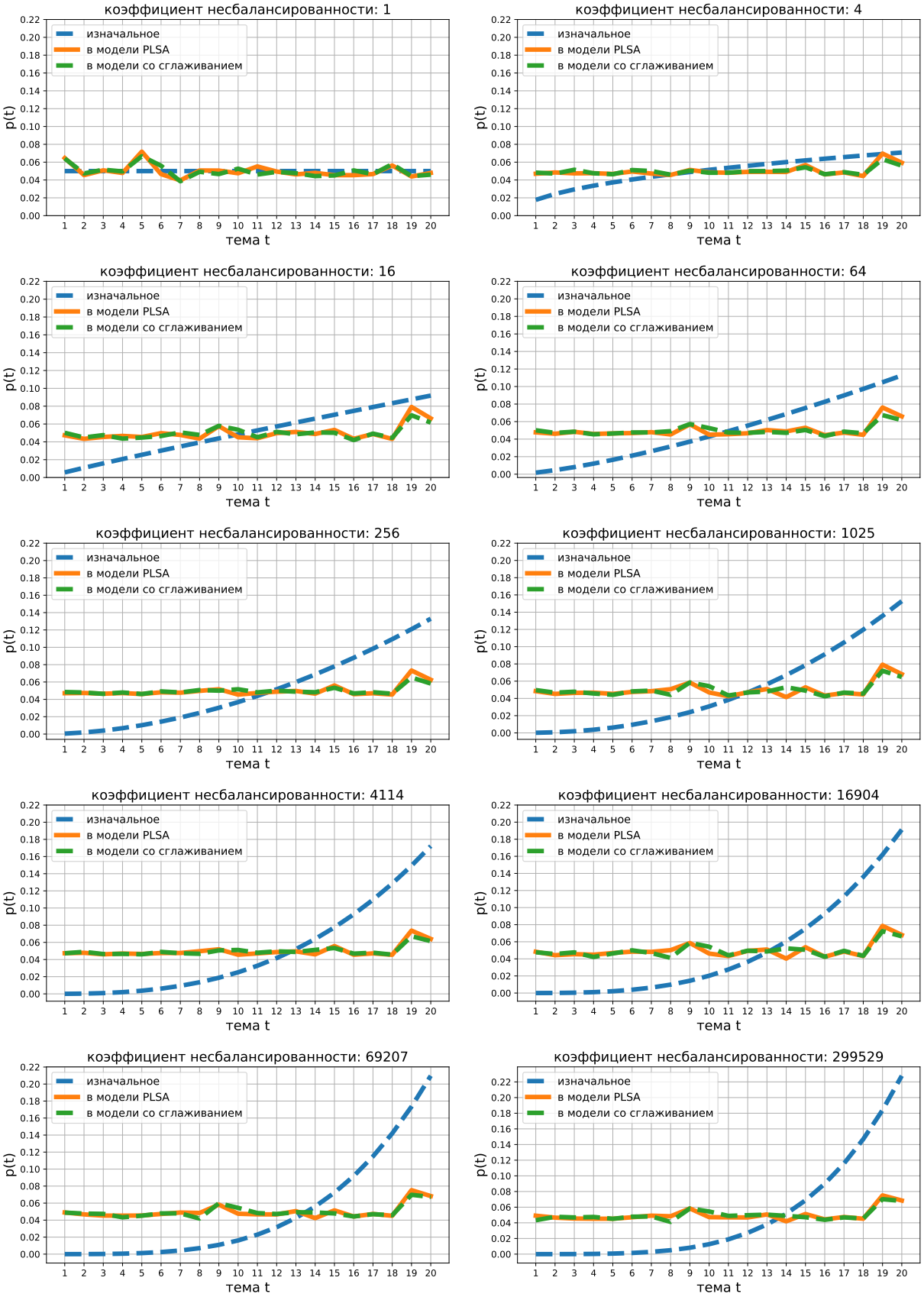


Рис. 10: распределение p(t), полученное моделью со сглаживанием, новые и исходные темы сопоставлены венгерским алгоритмом.

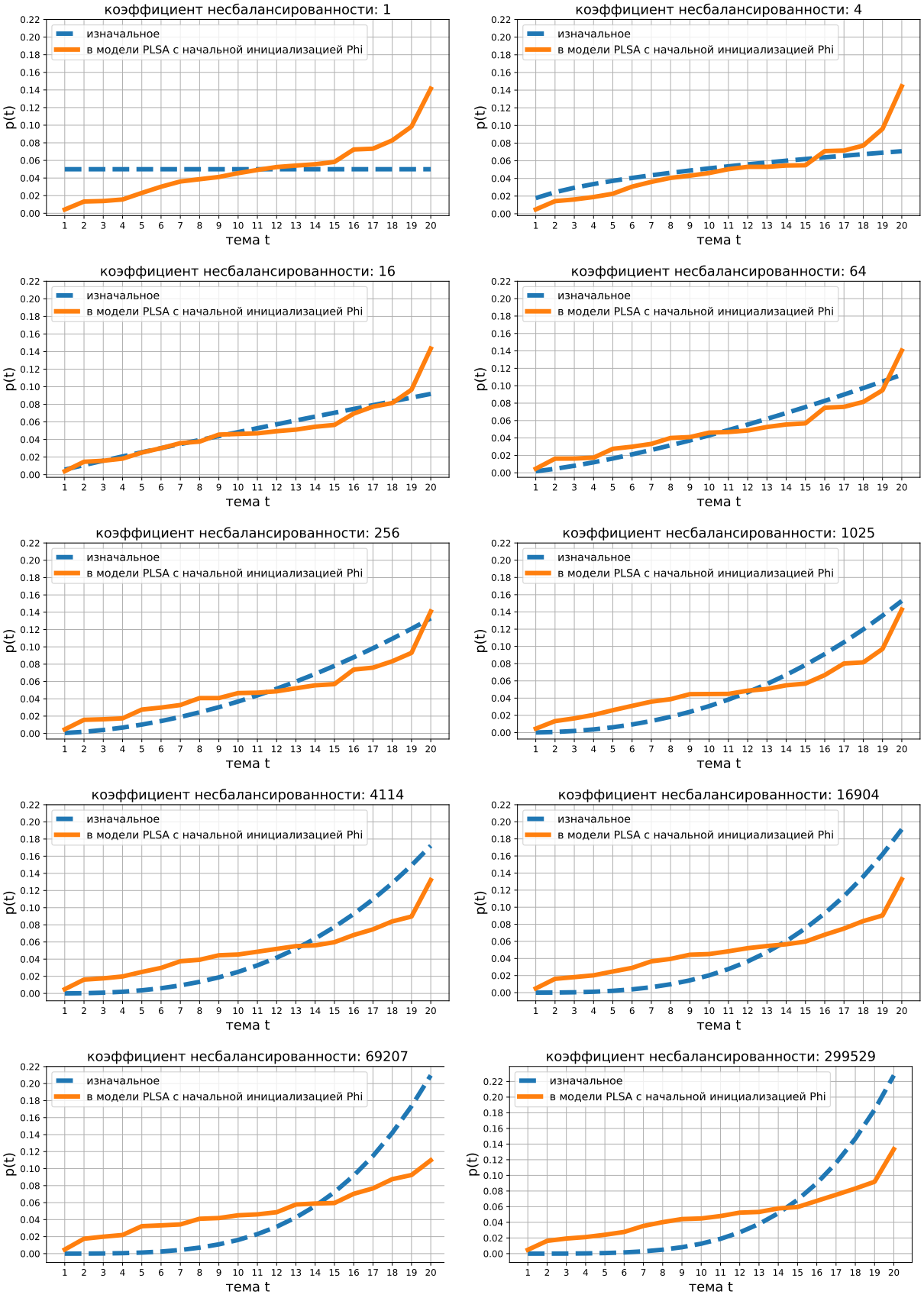


Рис. 11: распределение p(t), полученное моделью с идеальной инициализацией, без сопоставления тем.

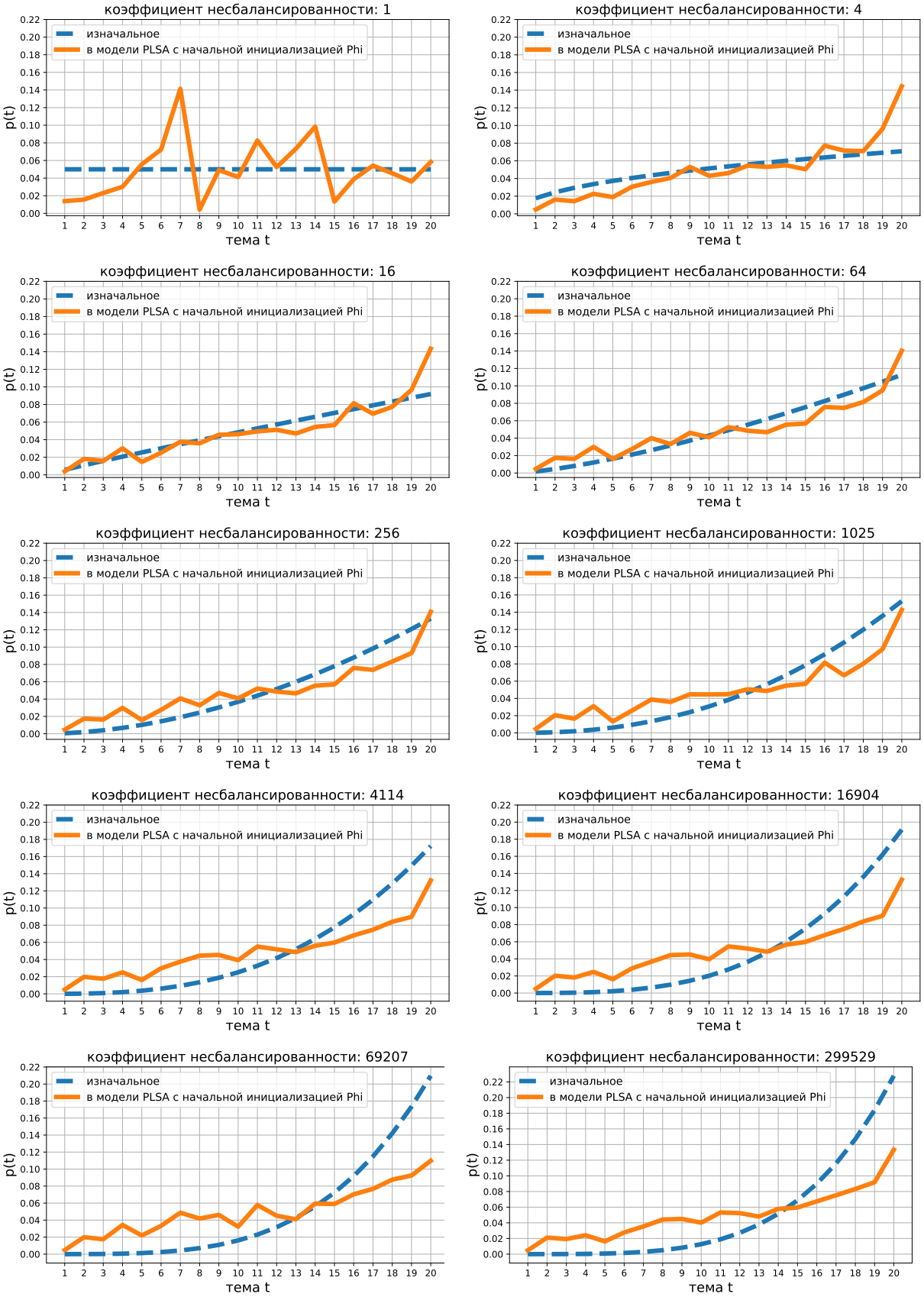


Рис. 12: распределение p(t), полученное моделью с идеальной инициализацией, новые и исходные темы сопоставлены венгерским алгоритмом.

Список литературы

[1] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '99. New York, NY, USA: ACM, 1999. Pp. 50– 57. http://doi.acm.org/10.1145/312624.312649.

[2] Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // J. Mach. Learn. Res. 2003. . Vol. 3. Pp. 993–1022. http://dl.acm.org/citation.cfm?id= 944919.944937.

[3] Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. Vol. 455. 2014.

[4] Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // Analysis of Images, Social networks and Texts. 2014.

[5] Tan Y., Ou Z. Topic-weak-correlated latent dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.

[6] K. Vorontsov, O. Frei, M. Apishev., P. Romov, M. Dudarenko. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // The 4th International Conference on Analysis of Images, Social Networks, and Texts (to appear), 2015.

[7] Hanna M. Wallach, David Mimno, Andrew McCallum: Rethinking LDA: Why Priors Matter // Neural Information Processing Systems, 2009