2.2 Марковский момент остановки агломеративного процесса кластеризации

После завершения агломеративного алгоритма кластеризации мы имеем множество минимальных расстояний $F_1, F_2, F_3, ..., F_{n-1}$, при этом для иерархических агломеративных алгоритмов (за исключением центроидного), числовые значения F_i монотонно возрастают: $0 \leq F_1 \leq F_2 \leq ... \leq F_{n-1}$ [26], [27].

При объединении «близких» кластеров, численные значения элементов множества минимальных расстояний возрастают медленно и их монотонное изменение почти линейно (если в случае центроидного метода $F_{i-1} > F_i$, то F_i заменяется на F_{i-1}). Так продолжается до тех пор, пока объединяемые кластеры достаточно близки. В момент, когда происходит объединение уже сформированных кластеров значение минимального расстояния резко возрастает. В этом случае для определения момента завершения процесса агломеративной кластеризации можно применить «метод локтя» и, соответственно, на рассматриваемом шаге абсцисса точки изгиба графика последовательности минимальных расстояний (обозначим её буквой k) совпадёт с предпочтительным количеством кластеров. Основная концепция аналитического обобщения «метода локтя» заключается в построении квадратичных форм аппроксимационно-оценочных критериев.

Обозначим набор $y_1, y_2, ..., y_k$ - множество тренда, полученное преобразованием $y_i = F_i + q \cdot i$, где q - коэффициент тренда, i - итерация в агломеративном процессе кластеризации. Рассматривается квадратическая погрешность линейной аппроксимации по m узлам:

$$\delta_l^2(m) = \sum_{i=0}^{m-1} (a \cdot i + b - y_i)^2$$

И неполная параболическая аппроксимация по m узлам:

$$\delta_q^2(m) = \sum_{i=0}^{m-1} (a \cdot i^2 + d - y_i)^2$$

Положим, что m = 4. Обозначим рассматриваемые на некоторой итерации узлы: $\hat{y_0}, \hat{y_1}, \hat{y_2}, \hat{y_3}$

Тогда, в качестве критерия используем разность двух погрешностей:

$$\delta^2(4_0) = \delta_l^2(4_0) - \delta_q^2(4_0) = \frac{1}{245}(19\hat{y_1}^2 - 11\hat{y_2}^2 + 41\hat{y_3}^2 + 12\hat{y_1}\hat{y_2} - 64\hat{y_1}\hat{y_3} - 46\hat{y_2}\hat{y_3})$$

Соответственно, характер возрастания числовой последовательности y_n изменится с линейного на параболический на k-ой итерации, если для узлов y_k, y_{k-1}, y_{k-2}, y_{k-3} справедливо неравенство $\delta > 0$, а для узлов y_{k-1}, y_{k-2}, y_{k-3}, y_{k-4} справедливо неравенство $\delta \leq 0$.

2.3 Применение методов к данным

Метод «AgglomerativeClustering», представленный в библиотеке «scikit-learn», в качестве параметров может принимать число кластеров и тип связи. В рассмотренной статье было получено 4 кластера. Будем предполагать, что это и будет оптимальным числом. В результате его применения получаем метки кластеров. В случае, когда параметр не указан - результат будет представлен в виде двух кластеров, что не представляет ценности для исследователя.

Результат после применения метода одиночной связи: кластеры размерами $(1, 1, 2, 78)$. Посмотрим на получившуюся дендограмму разбиения 2.

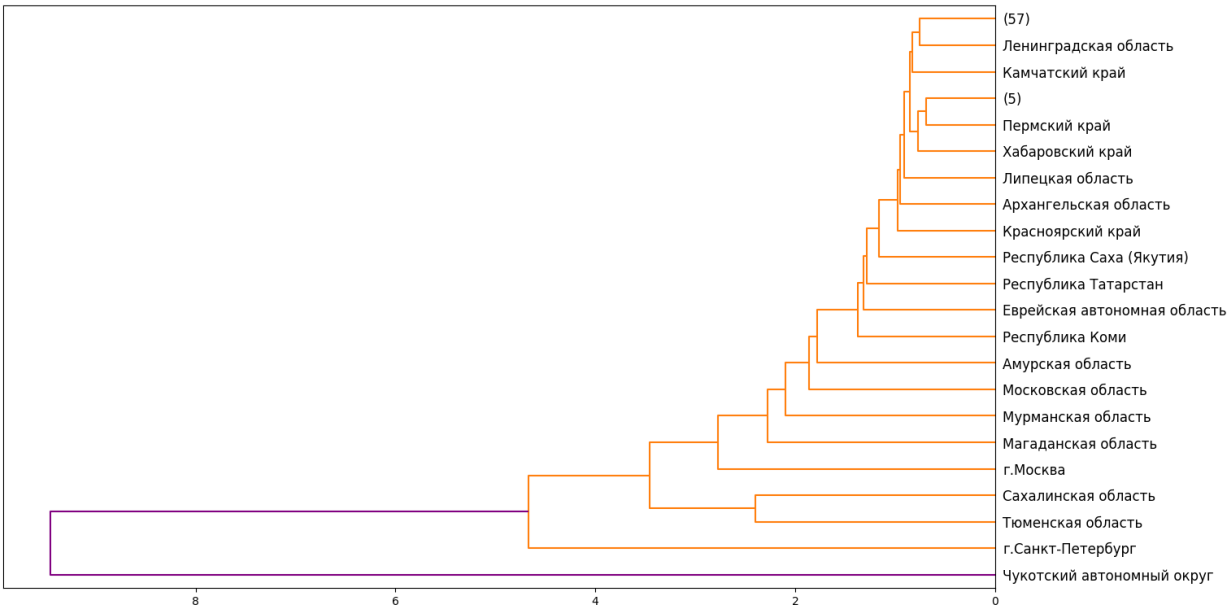

Рис. 2: Дендограмма (одиночная связь)