问题1-改进

改进：使用模型改进数据重新聚类

改进前：漏水量 F_l 直接取 2-5 点用户用水量的最小值 F_{min}

改进后：漏水量服从一个与 2-5 点用户用水量的分布，如下：

$$F_l = \begin{cases} F_{min} & ,if\ F_{min} \leq 25 \\ (F_{min} - 25) + 0.7F_{min} & ,if\ 25 < F_{min} \leq 30 \\ (F_{min} - 25) + 0.5F_{min} & ,else \end{cases}$$

In [1]:

```python
import numpy as np
import pandas as pd
import cufflinks as cf

import scipy
import scipy.cluster.hierarchy as sch
from sklearn.metrics import *

import plotly
import plotly.express as px
import plotly.graph_objects as go
import plotly.figure_factory as ff

import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False

from IPython.display import HTML
from IPython.core.interactiveshell import InteractiveShell
# InteractiveShell.ast_node_interactivity = 'all'
InteractiveShell.ast_node_interactivity = 'last'

import pylatex
import latexify
```

层次聚类 (剔除 5-28 号)

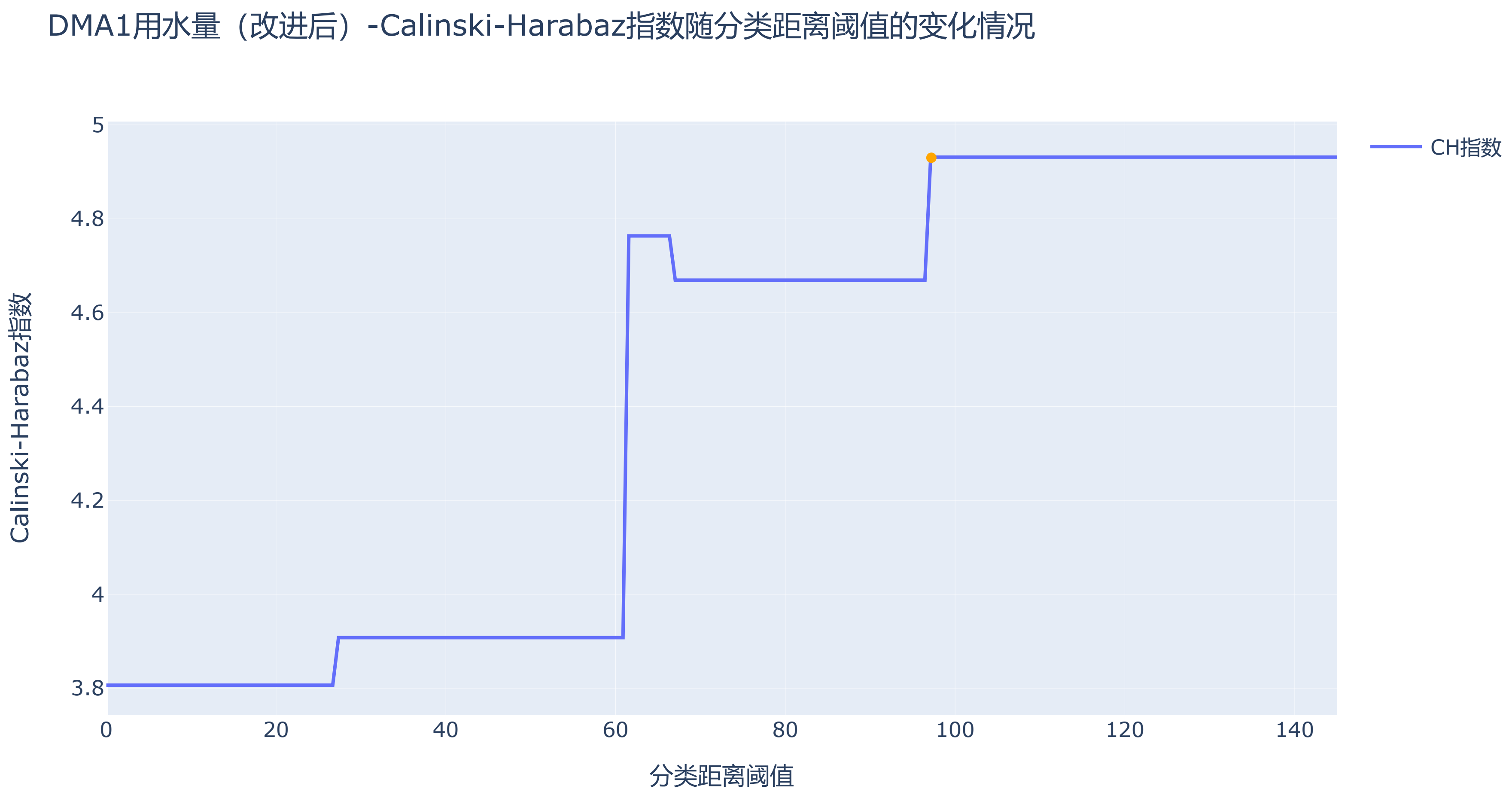
DMA 1 日期 用水量聚类

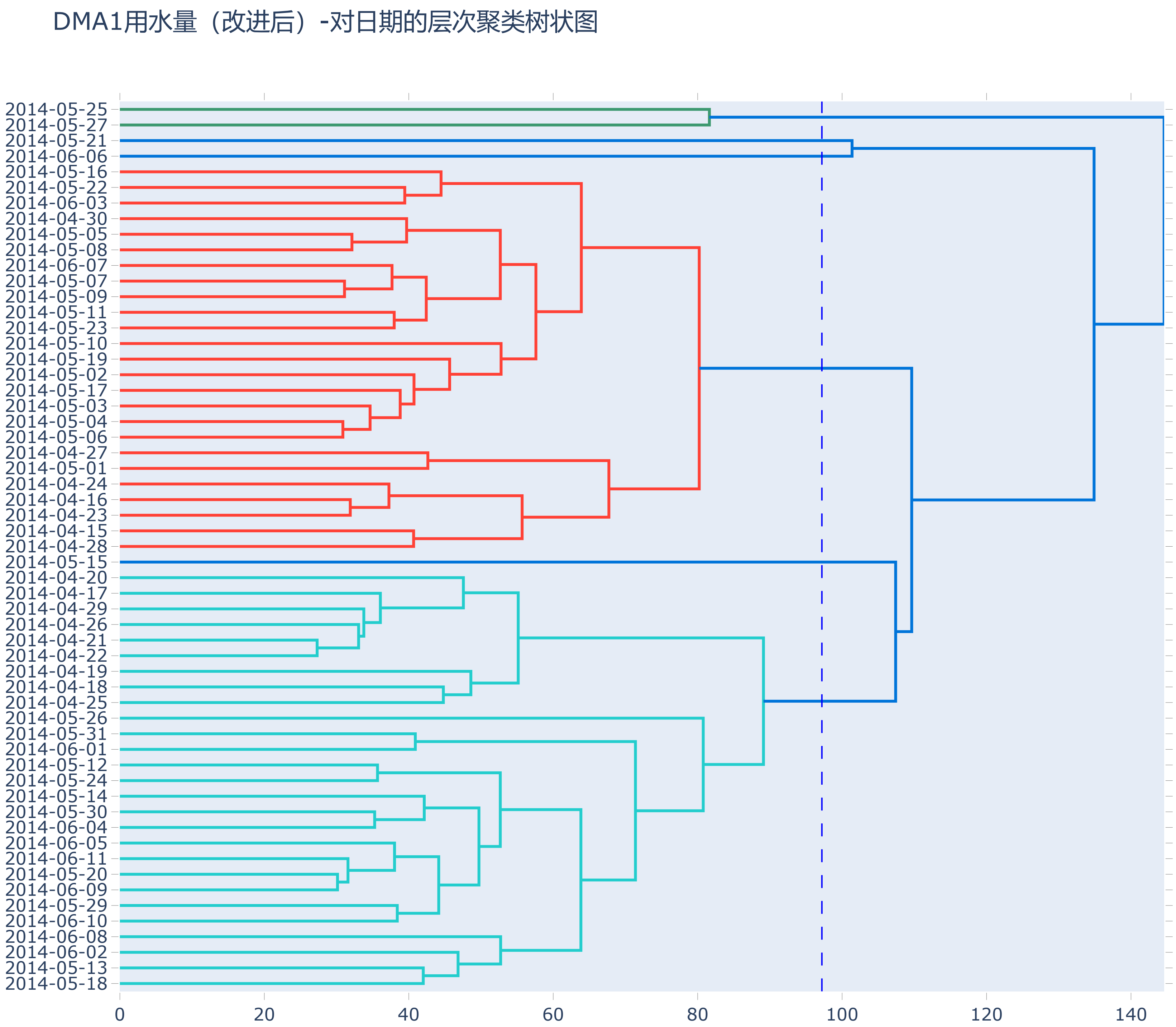
In [2]:

```
# DMA1 data
user_DMA1 = pd.read_excel("模型改进数据.xlsx", sheet_name='DMA1的用户用水量', index_col=0)
user_DMA1 = pd.concat([user_DMA1.iloc[:43, :], user_DMA1.iloc[44:, :]])  # 剔除 5-28
index = list(user_DMA1.index.strftime("%Y-%m-%d"))
columns = list(user_DMA1.columns)
```

In [3]:

```
# InteractiveShell.ast_node_interactivity = 'all'
InteractiveShell.ast_node_interactivity = 'last'

dis_arr = np.array(user_DMA1)
disMat = sch.distance.pdist(dis_arr, 'euclidean')
Z = sch.linkage(disMat)

cluster = sch.fcluster(Z, 1, 'inconsistent')
ch_score = []
b = 1.06140
t = np.linspace(0, b, int(200*(b)+1))
tt = np.linspace(0, 145, int(200*(b)+1))
for d in t:
    cluster = sch.fcluster(Z, d, 'inconsistent')
    s = calinski_harabasz_score(user_DMA1, cluster)
    ch_score.insert(0, s)
    ch_score.insert(0, ch_score[0])
    ch_score.pop()
# len(set(sch.fcluster(Z, 0.88, 'inconsistent')))
trace = go.Scatter(x=tt, y=ch_score, mode='lines', name='CH指数')
fig = go.Figure(data=trace)
fig.update_layout(
    width=910,
    xaxis=dict(title='分类距离阈值'),
    yaxis=dict(title='Calinski-Harabaz指数'),
    title_text="DMA1用水量（改进后）-Calinski-Harabaz指数随分类距离阈值的变化情况",
)
fig.add_trace(go.Scatter(
    x=[97.2], y=[4.93],
    line=dict(color='orange', width=5),
    showlegend=False,
))
fig.write_image('./img/svg/DMA1用水量（改进后）-Calinski-Harabaz指数随分类距离阈值的变化情况.svg')
fig.show()

fig = ff.create_dendrogram(user_DMA1, orientation='left', labels=index, )
fig.update_layout(
    width=900,
    height=800,
    yaxis=dict(range=[-570, 0]),
    title_text='DMA1用水量（改进后）-对日期的层次聚类树状图',
)
fig.add_trace(go.Scatter(
    x=[97.2] * len(ch_score),
    y=np.linspace(-570, 0, len(ch_score)),
    mode='lines',
    line=dict(color='blue', width=1, dash='dash'),
))
fig.write_image('./img/svg/DMA1用水量（改进后）-对日期进行层次聚类结果.svg')
fig.show()
```



DMA1用水量（改进后）-对日期的层次聚类树状图
2014-05-25
2014-05-27
2014-05-21
2014-06-06
2014-05-16
2014-05-22
2014-06-03
2014-04-30
2014-05-05
2014-05-08
2014-06-07
2014-05-07
2014-05-09
2014-05-11
2014-05-23
2014-05-10
2014-05-19
2014-05-02
2014-05-17
2014-05-03
2014-05-04
2014-05-06
2014-04-27
2014-05-01
2014-04-24
2014-04-16
2014-04-23
2014-04-15
2014-04-28
2014-05-15
2014-04-20
2014-04-17
2014-04-29
2014-04-26
2014-04-21
2014-04-22
2014-04-19
2014-04-18
2014-04-25
2014-05-26
2014-05-31
2014-06-01
2014-05-12
2014-05-24
2014-05-14
2014-05-30
2014-06-04
2014-06-05
2014-06-11
2014-05-20
2014-06-09
2014-05-29
2014-06-10
2014-06-08
2014-06-02
2014-05-13
2014-05-18
0
20
40
60
80
100
120
140

In [ ]:

```

```

DMA 2 日期 用水量聚类

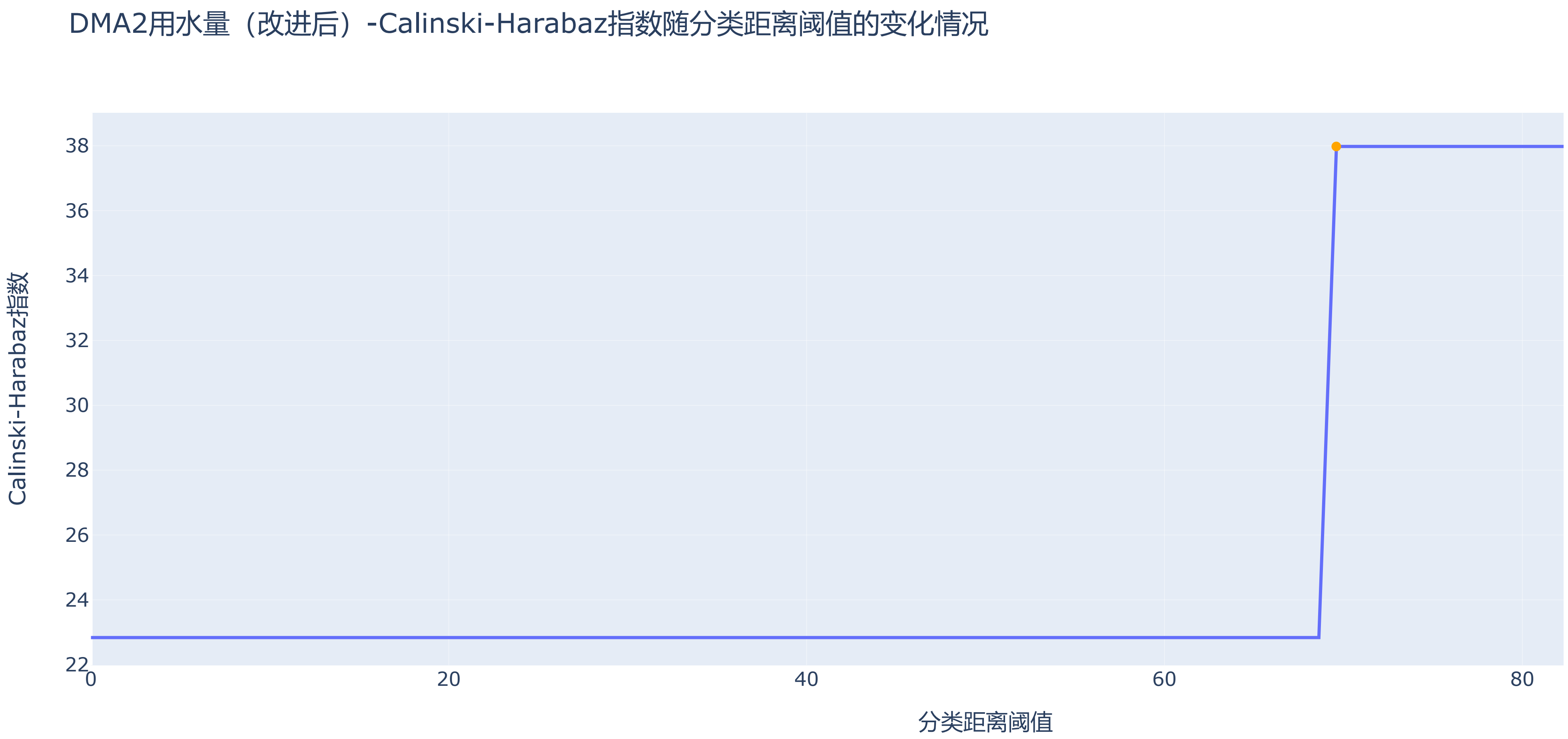
In [4]:

```
# DMA2 data
user_DMA2 = pd.read_excel("模型改进数据.xlsx", sheet_name='DMA2的用户用水量', index_col=0)
user_DMA2 = pd.concat([user_DMA2.iloc[:43, :], user_DMA2.iloc[44:, :]])  # 剔除 5-28
index = list(user_DMA2.index.strftime("%Y-%m-%d"))
columns = list(user_DMA2.columns)
```

In [5]:

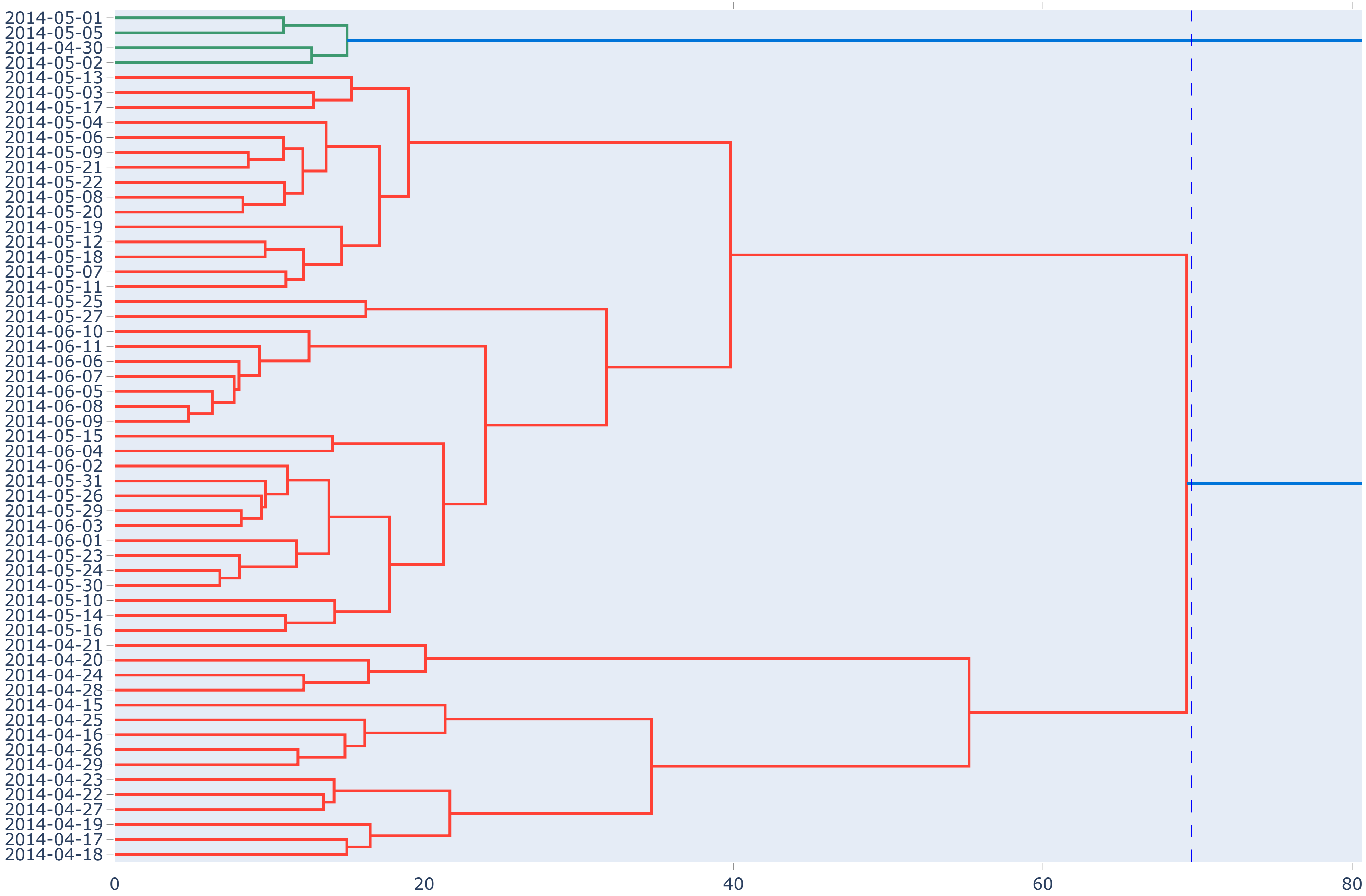
```
# InteractiveShell.ast_node_interactivity = 'all'
InteractiveShell.ast_node_interactivity = 'last'

dis_arr = np.array(user_DMA2)
disMat = sch.distance.pdist(dis_arr, 'euclidean')
Z = sch.linkage(disMat)
ch_score = []
b = 1.02
t = np.linspace(0, b, int(100*(b)+1))
tt = np.linspace(0, 100, int(100*(b)+1))
for d in t:
    cluster = sch.fcluster(Z, d, 'inconsistent')  # 聚类结果
    s = calinski_harabasz_score(user_DMA2, cluster)
    ch_score.append(s)
# len(set(sch.fcluster(Z, 0.97, 'inconsistent')))
trace = go.Scatter(x=tt, y=ch_score, mode='lines', name='CH指数')
fig = go.Figure(data=trace)
fig.update_layout(
    width=1320,
    xaxis=dict(title='分类距离阈值'),
    yaxis=dict(title='Calinski-Harabaz指数'),
    title_text="DMA2用水量（改进后）-Calinski-Harabaz指数随分类距离阈值的变化情况",
)
fig.add_trace(go.Scatter(
    x=[69.6], y=[37.98],
    line=dict(color='orange', width=5),
    showlegend=False,
))
fig.write_image('./img/svg/DMA2用水量（改进后）-Calinski-Harabaz指数随分类距离阈值的变化情况.svg')
fig.show()


fig = ff.create_dendrogram(user_DMA2, orientation='left', labels=index)
```

```
fig.update_layout(
    width=1300,
    height=800,
    yaxis=dict(range=[-570, 0]),
    title_text='DMA2用水量（改进后）-对日期的层次聚类树状图',
)
fig.add_trace(go.Scatter(
    x=[69.6] * len(ch_score),
    y=np.linspace(-570, 0, len(ch_score)),
    mode='lines',
    line=dict(color='blue', width=1, dash='dash'),
))
fig.write_image('./img/svg/DMA2用水量（改进后）-对日期进行层次聚类结果.svg')
fig.show()
```



DMA2用水量（改进后）-Calinski-Harabaz指数随分类距离阈值的变化情况

DMA2用水量（改进后）-对日期的层次聚类树状图



In [ ]: