

# 数据集整理

## 1. REFIT 电气负载测量数据集

- 规模

Appliance Type	# of Uses	Consumption (kWh)	# of Appliances
Fridge Freezers	121,752	20,020	13
Fridge	53,163	1,310	7
Freezers	133,967	5,486	13
Washing Machines	6,865	3,994	21
Dishwashers	4,250	6,827	14
Tumble Dryers	2,372	4,210	10
Kettle	40,092	3,298	16
Microwave	12,946	1,208	16
Toaster	5,364	257	9
ICT Equipment	4,176	3,104	13
Television Site	11,274	5,995	21
Electric Heater	503	1,023	4
Bread Maker	206	56	1
Pond Pump	Continuous	282	1
Vivarium	Continuous	208	1

- 简介

此数据集包含智能电表读数和各个设备能耗。数据集是在英国的20栋房屋中用两年时间收集的。在数据收集过程中，住户们正在执行他们通常的例行公事。能耗取决于居住者人数、电器数量和卧室数量。

- 下载链接

[https://www.researchgate.net/publication/312267812\\_An\\_electrical\\_load\\_measurements\\_dataset\\_of\\_United\\_Kingdom\\_households\\_from\\_a\\_two-year\\_longitudinal\\_study](https://www.researchgate.net/publication/312267812_An_electrical_load_measurements_dataset_of_United_Kingdom_households_from_a_two-year_longitudinal_study)

- 数据示例

House	Dates	Metered [kWh]
8	29/09/2014–15/10/2014	226
8	15/10/2014–27/01/2015	1785
8	27/01/2015–05/03/2015	657
10	15/10/2014–24/03/2015	2799
13	04/10/2014–26/11/2014	640
17	13/11/2014–02/12/2014	178
18	15/10/2014–18/11/2014	333
18	18/11/2014–16/12/2014	316
19	02/12/2014–11/12/2014	79

Metered represents the difference in readings between the two dates which was recorded by the utility installed meter for the house.

## 2. 电力负荷图2011–2014数据集

- 规模

Data Set Characteristics:	Time-Series	Number of Instances:	370	Area:	Computer
Attribute Characteristics:	Real	Number of Attributes:	140256	Date Donated	2015-03-13
Associated Tasks:	Regression, Clustering	Missing Values?	N/A	Number of Web Hits:	79138

- 简介

此数据集是从2011年到2014年每15分钟记录一次用电量（单位为kWh），获取得到的370个用户的用电量实例，并将其转化为小时用电量展示，其中2011年未开通服务的用户的数据用0填充。

- 下载链接

<https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014>

- 数据格式

数据集使用csv格式，保存为txt。第一列以字符串形式显示日期和时间，格式为'yyyy-mm-dd hh: mm: ss'。其他列显示浮点值，其消耗量以kW为单位。

## 3. Numenta异常基准 (NAB) 数据集

- 简介

NAB是评估流实时应用中异常检测算法的新颖基准。它由50多个带有标签的真实世界和人工时间序列数据文件以及为实时应用程序设计的新颖评分机制组成。其中包含的工具可以为模型使用本数据集中数据进行实验的结果进行统一的量化与评估，开源项目的结果将可以公示在Scoreboard中进行比较与展示（得分从0到100标志着模型异常检测的结果从坏到好）。

- 下载链接

<https://github.com/numenta/NAB>

- 数据内容说明

数据包括真实数据与人工生成的数据，数据均有序、带有时间戳且包括异常，数据的来源涵盖了AWS服务器指标、网站点击率、交通数据，Twitter数据等。

## 4. 加利福尼亚交通运输数据集 (PeMS数据集)

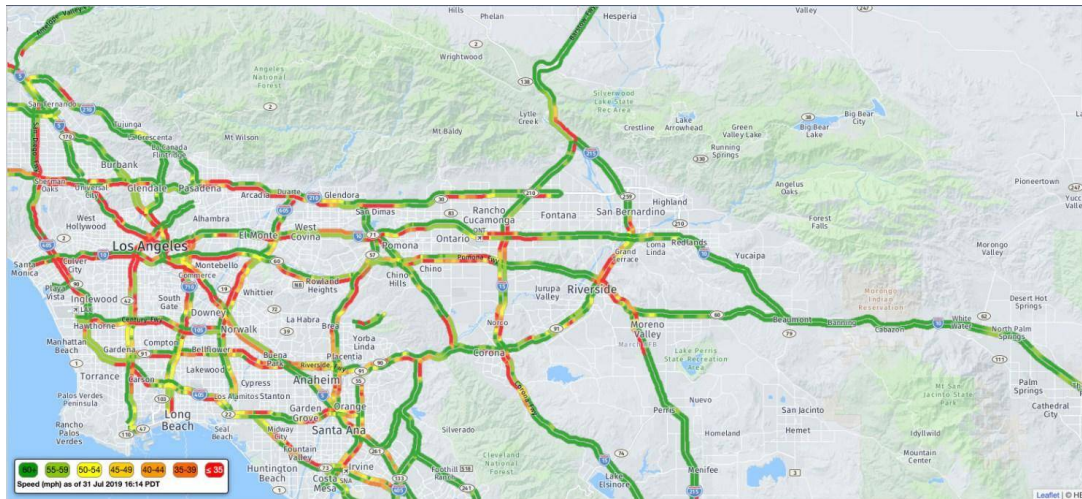
- 简介

数据集中包含的交通数据是从39,000多个单独的探测器中实时收集的。这些传感器跨越了加利福尼亚州所有主要都会区的高速公路系统。

- 下载链接

<http://pems.dot.ca.gov/> （需要注册下载）

- 数据展示



- 数据内容说明

数据内容包括交通探测器信息、突发事件、车道封闭、流量计数、车辆分类、道路清单等内容。

## 5. Yahoo's Webscope S5 数据集

- 简介

该数据集由带有标记异常点的真实时间序列和合成时间序列组成，用于测试异常值和变化点等各类异常值的检测精度。其中，真实的数据集由时间序列组成，代表了各种雅虎服务的指标；合成的数据集则具有变化趋势、噪声和季节性的时间序列组成。

- 下载链接

<https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70>

- 数据展示

目前已经向Yahoo提供数据集的需求申请，还没得到下载许可

## 6. 2018 AIOps's KPI-Anomaly-Detection 数据集

- 简介

该数据集由来自互联网公司的许多真实场景的KPI（关键绩效指数）时间序列数据和ground truth标签组成。KPI分为两大类：服务KPI和机器KPI。服务kpi是反映Web服务的大小和质量的性能指标，如页面响应时

间、页面视图和连接错误数量。机器kpi是反映机器(服务器、路由器、交换机)运行状况的性能指标，如CPU利用率、内存利用率、磁盘IO、网卡吞吐量等。

该数据集包含了训练数据集和测试数据集。训练数据集中的每一组数据包括KPI ID、Timestamp、Value和Label四个属性；测试数据集中的每一组数据包括KPI ID、Timestamp和Value三个属性。

- 下载链接

<https://github.com/rob-med/awesome-TS-anomaly-detection>

- 数据示例

- 训练数据集

KPI ID	Timestamp	Value	Label
0	1503831000	10.8	0
0	1503831060	12.3	1
...	...	...	...

- 测试数据集

KPI ID	Timestamp	Value
0	1503831000	10.8
0	1503831060	12.3
...	...	...