Data Modeling Project for MATH 7241

October 2020

Project

$\boxed{\text{Due date}}$: Tuesday December 1, week after Thanksgiving.

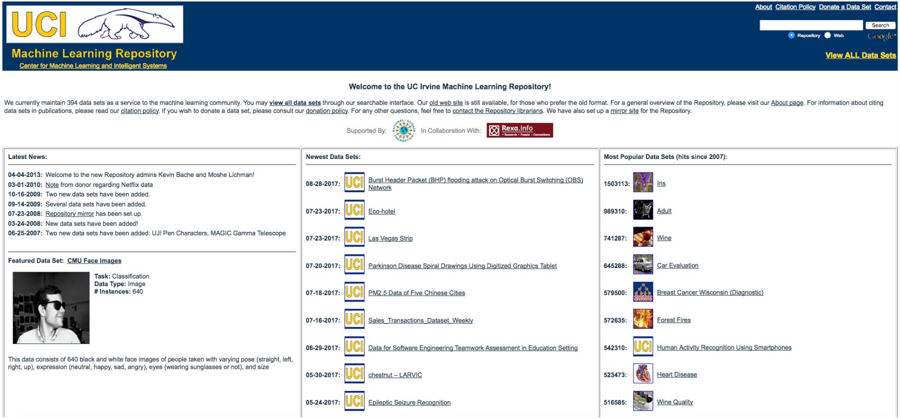
Groups: if you wish you may collaborate in a group of at most two people. In this case each member of the group must make a substantial contribution to the project, and your project report must describe which contributions were made by each member.

Contents

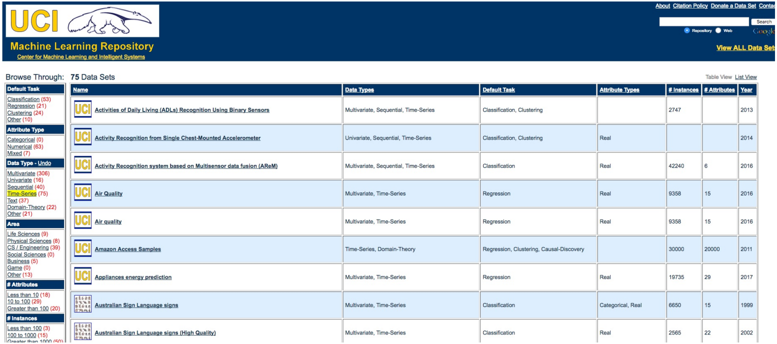
$\boxed{\text{Step 1:}}$ find a good data set!

Recommended source: UCI archive. Search for Time Series.

https://archive.ics.uci.edu/ml/index.php


About Citation Policy Donate a Data Set Contact
UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
Search
Repository Web
View ALL Data Sets
Welcome to the UC Irvine Machine Learning Repository!
We currently maintain 394 data sets as a service to the machine learning community. You may view all data sets through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Repository.
Supported By: In Collaboration With: Rexa.info
Latest News:
04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: Note from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
07-23-2008: Repository mirror has been set up.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope
Featured Data Set: CMU Face Images
Task: Classification
Data Type: Image
Instances: 640
This data consists of 640 black and white face images of people taken with varying pose (straight, left, right, up), expression (neutral, happy, sad, angry), eyes (wearing sunglasses or not), and size
Newest Data Sets:
08-28-2017: Burst Header Packet (BHP) flooding attack on Optical Burst Switching (OBS) Network
07-23-2017: Eco-hotel
07-23-2017: Las Vegas Strip
07-20-2017: Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet
07-18-2017: PM2.5 Data of Five Chinese Cities
07-16-2017: Sales_Transactions_Dataset_Weekly
06-29-2017: Data for Software Engineering Teamwork Assessment in Education Setting
05-30-2017: chestnut – LARVIC
05-24-2017: Epileptic Seizure Recognition
Most Popular Data Sets (hits since 2007):
1503113: Iris
989310: Adult
741287: Wine
645288: Car Evaluation
579500: Breast Cancer Wisconsin (Diagnostic)
572635: Forest Fires
542310: Human Activity Recognition Using Smartphones
523473: Heart Disease
516585: Wine Quality

Search for a time series:



Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Activities of Daily Living (ADLs) Recognition Using Binary Sensors	Multivariate, Sequential, Time-Series	Classification, Clustering		2747		2013
Activity Recognition from Single Chest-Mounted Accelerometer	Univariate, Sequential, Time-Series	Classification, Clustering	Real			2014
Activity Recognition system based on Multisensor data fusion (AReM)	Multivariate, Sequential, Time-Series	Classification	Real	42240	6	2016
Air Quality	Multivariate, Time-Series	Regression	Real	9358	15	2016
Air quality	Multivariate, Time-Series	Regression	Real	9358	15	2016
Amazon Access Samples	Time-Series, Domain-Theory	Regression, Clustering, Causal-Discovery		30000	20000	2011
Appliances energy prediction	Multivariate, Time-Series	Regression	Real	19735	29	2017
Australian Sign Language signs	Multivariate, Time-Series	Classification	Categorical, Real	6650	15	1999
Australian Sign Language signs (High Quality)	Multivariate, Time-Series	Classification	Real	2565	22	2002

Select one time series:

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Air Quality Data Set

Download: Data Folder, Data Set Description

Abstract: Contains the responses of a gas multisensor device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	9358	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	15	**Date Donated**	2016-03-23
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	131942

Source:

Saverio De Vito (saverio.devito '@' enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development

Data Set Information:

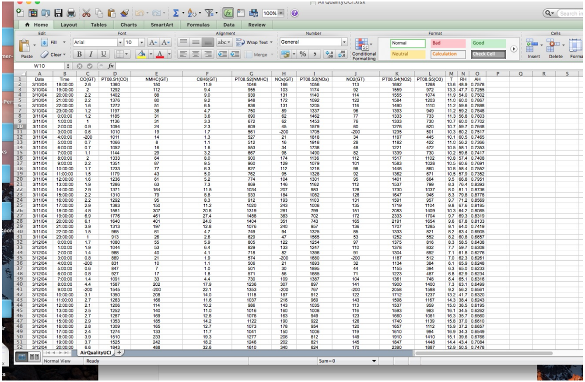
The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.
This dataset can be used exclusively for research purposes. Commercial purposes are fully excluded.

Attribute Information:

0 Date (DD/MM/YYYY)
1 Time (HH.MM.SS)
2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7 True hourly averaged NOx concentration in ppb (reference analyzer)
8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)

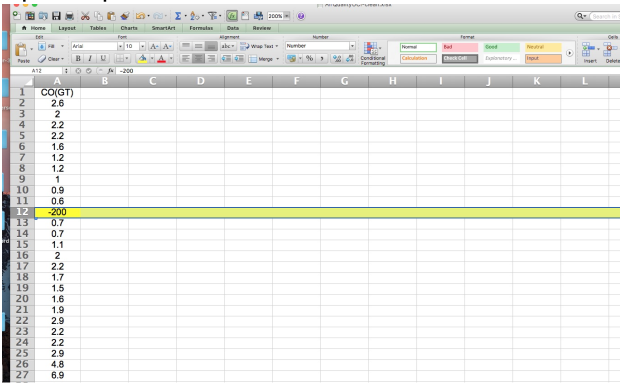
$\boxed{\text{Step 2:}}$ download and clean the data set to prepare for modeling. For example, remove unnecessary data, correct ‘error’ entries etc

Raw data set:



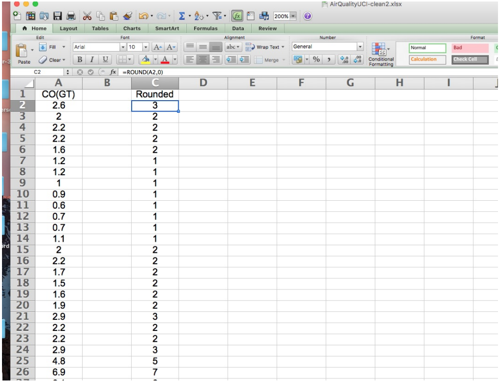
Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
3/10/04	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
3/10/04	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
3/10/04	20:00:00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
3/10/04	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
3/10/04	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
3/10/04	23:00:00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848
3/11/04	0:00:00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603
3/11/04	1:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0	0.7702
3/11/04	2:00:00	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7	0.7648
3/11/04	3:00:00	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2	0.7517
3/11/04	4:00:00	-200	1011	14	1.3	527	21	1818	34	1197	445	10.1	60.5	0.7465
3/11/04	5:00:00	0.7	1066	8	1.1	512	16	1918	28	1182	422	11.0	56.2	0.7366
3/11/04	6:00:00	0.7	1052	16	1.6	553	34	1738	48	1221	472	10.5	58.1	0.7353
3/11/04	7:00:00	1.1	1144	29	3.2	667	98	1490	82	1339	730	10.2	59.6	0.7417
3/11/04	8:00:00	2	1333	64	8.0	900	174	1136	112	1517	1102	10.8	57.4	0.7408
3/11/04	9:00:00	2.2	1351	87	9.5	960	129	1079	101	1583	1028	10.5	60.6	0.7691
3/11/04	10:00:00	1.7	1233	77	6.3	827	112	1218	98	1446	860	10.8	58.4	0.7552
3/11/04	11:00:00	1.5	1179	43	5.0	762	95	1328	92	1362	671	10.5	57.9	0.7352
3/11/04	12:00:00	1.6	1236	61	5.2	774	104	1301	95	1401	664	9.5	66.8	0.7951
3/11/04	13:00:00	1.9	1286	63	7.3	869	146	1162	112	1537	799	8.3	76.4	0.8393
3/11/04	14:00:00	2.9	1371	164	11.5	1034	207	983	128	1730	1037	8.0	81.1	0.8736
3/11/04	15:00:00	2.2	1310	79	8.8	933	184	1082	126	1647	946	8.3	79.8	0.8778
3/11/04	16:00:00	2.2	1292	95	8.3	912	193	1103	131	1591	957	9.7	71.2	0.8569
3/11/04	17:00:00	2.9	1383	150	11.2	1020	243	1008	135	1719	1104	9.8	67.6	0.8185
3/11/04	18:00:00	4.8	1581	307	20.8	1319	281	799	151	2083	1409	10.3	64.2	0.8065
3/11/04	19:00:00	6.9	1776	461	27.4	1488	383	702	172	2333	1704	9.7	69.3	0.8319
3/11/04	20:00:00	6.1	1640	401	24.0	1404	351	743	165	2191	1654	9.6	67.8	0.8133
3/11/04	21:00:00	3.9	1313	197	12.8	1076	240	957	136	1707	1285	9.1	64.0	0.7419
3/11/04	22:00:00	1.5	965	61	4.7	749	94	1325	85	1333	821	8.2	63.4	0.6905
3/11/04	23:00:00	1	913	26	2.6	629	47	1565	53	1252	552	8.2	60.8	0.6657
3/12/04	0:00:00	1.7	1080	55	5.9	805	122	1254	97	1375	816	8.3	58.5	0.6438
3/12/04	1:00:00	1.9	1044	53	6.4	829	133	1247	110	1378	832	7.7	59.7	0.6308
3/12/04	2:00:00	1.4	988	40	4.1	718	82	1396	91	1304	692	7.1	61.8	0.6276
3/12/04	3:00:00	0.8	889	21	1.9	574	-200	1680	-200	1187	512	7.0	62.3	0.6261
3/12/04	4:00:00	-200	831	10	1.1	506	21	1893	32	1134	384	6.1	65.9	0.6248
3/12/04	5:00:00	0.6	847	7	1.0	501	30	1895	44	1155	394	6.3	65.0	0.6233
3/12/04	6:00:00	0.8	927	17	1.8	571	56	1685	71	1223	487	6.8	62.9	0.6234
3/12/04	7:00:00	1.4	1091	33	4.4	730	109	1387	104	1361	748	6.4	65.1	0.6316
3/12/04	8:00:00	4.4	1587	202	17.9	1236	307	897	141	1900	1400	7.3	63.1	0.6499
3/12/04	9:00:00	-200	1545	-200	22.1	1353	-200	767	-200	2058	1588	9.2	56.2	0.6561
3/12/04	10:00:00	3.1	1350	208	14.0	1118	187	912	122	1712	1237	13.2	41.7	0.6320
3/12/04	11:00:00	2.7	1263	166	11.6	1037	216	969	143	1598	1167	14.3	38.4	0.6243
3/12/04	12:00:00	2.1	1206	114	10.2	986	143	1035	113	1537	959	15.0	36.5	0.6195
3/12/04	13:00:00	2.5	1252	140	11.0	1016	160	1008	116	1593	983	16.1	34.5	0.6262
3/12/04	14:00:00	2.7	1287	169	12.8	1078	163	949	123	1660	1061	16.3	35.7	0.6560
3/12/04	15:00:00	2.9	1353	185	14.2	1122	190	922	126	1740	1139	15.8	37.0	0.6610
3/12/04	16:00:00	2.8	1309	165	12.7	1073	178	954	120	1657	1112	15.9	37.2	0.6657
3/12/04	17:00:00	2.4	1274	133	11.7	1041	150	1006	119	1610	994	16.9	34.3	0.6549
3/12/04	18:00:00	3.9	1510	233	19.3	1277	206	812	149	1910	1410	15.1	39.6	0.6766
3/12/04	19:00:00	3.7	1525	242	18.2	1246	202	821	145	1847	1448	14.4	43.4	0.7084
3/12/04	20:00:00	6.6	1843	488	32.6	1610	340	624	170	2390	1887	12.9	50.5	0.7478

Clean it up: keep one column, remove error entries:



$\boxed{\text{Step 3:}}$ map your data into a Markov chain. To do this, you must choose the states for your model. Each entry in the time series should map into a unique state.

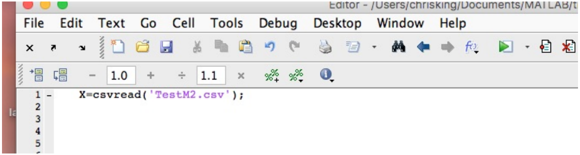
In this example, we choose a 9-state model, with the states $\{1, 2, \ldots, 9\}$. Map each entry into a state by rounding to the nearest integer.



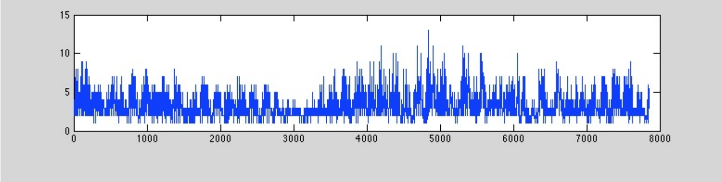
	A	B	C
1	CO(GT)		Rounded
2	2.6		3
3	2		2
4	2.2		2
5	2.2		2
6	1.6		2
7	1.2		1
8	1.2		1
9	1		1
10	0.9		1
11	0.6		1
12	0.7		1
13	0.7		1
14	1.1		1
15	2		2
16	2.2		2
17	1.7		2
18	1.5		2
19	1.6		2
20	1.9		2
21	2.9		3
22	2.2		2
23	2.2		2
24	2.9		3
25	4.8		5
26	6.9		7

$\boxed{\text{Step 4:}}$ transfer the time series to a platform for analysis, eg Excel, Matlab, R etc.

For example, here we import the spreadsheet into Matlab:

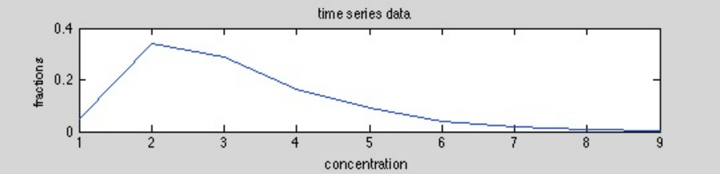


and here is the complete time series:



$\boxed{\text{Step 5:}}$ compute the occupation frequencies for each state, and turn this into a probability distribution. This is the empirical distribution of your chain.

Empirical distribution from time series: fraction of time spent in each state



$\boxed{\text{Step 6:}}$ compute the frequencies of jumps between each pair of states. Divide by the occupation frequency at each state so that the total jump probability out of each state is 1. This is your transition matrix.

Transition matrix from time series:

Command Window

File Edit Debug Desktop Window Help

New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
Trans2 =

  Columns 1 through 7

    0.6393    0.3532    0.0050    0.0025         0         0         0
    0.0540    0.7504    0.1507    0.0317    0.0110    0.0015    0.0008
    0.0009    0.2129    0.5973    0.1443    0.0356    0.0058    0.0018
         0    0.0297    0.3143    0.4425    0.1572    0.0414    0.0133
         0    0.0029    0.1225    0.3098    0.3429    0.1614    0.0490
         0    0.0031    0.0404    0.2112    0.2888    0.2360    0.1553
         0         0    0.0132    0.1060    0.2384    0.2914    0.1589
         0         0    0.0161    0.0484    0.2258    0.2097    0.1935
         0         0         0    0.0323    0.0323    0.1935    0.1290

  Columns 8 through 9

         0         0
         0         0
    0.0009         0
    0.0016         0
    0.0072    0.0014
    0.0404    0.0248
    0.1258    0.0596
    0.1774    0.0806
    0.1935    0.1935

fx >>
```

$\boxed{\text{Step 7:}}$ find the stationary vector of your transition matrix. In case it is not unique, find all stationary vectors.

Stationary distribution of Markov chain:

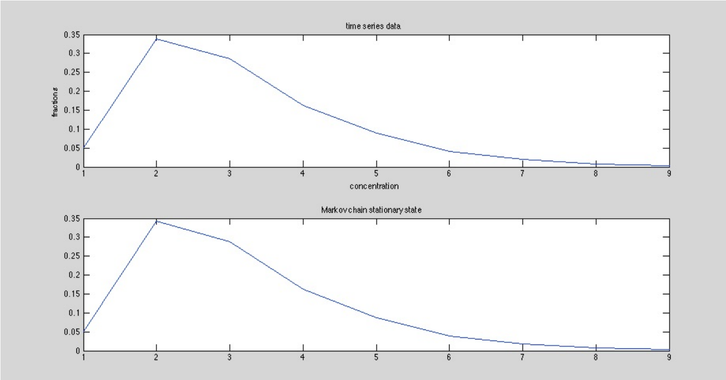
```
ans =

    0.0522    0.3425    0.2879    0.1623    0.0869    0.0397    0.0180    0.0069    0.0034

fx >>
```

$\boxed{\text{Step 8:}}$ compare the empirical distribution of the data set and the stationary distribution of your chain. Note any similarities!

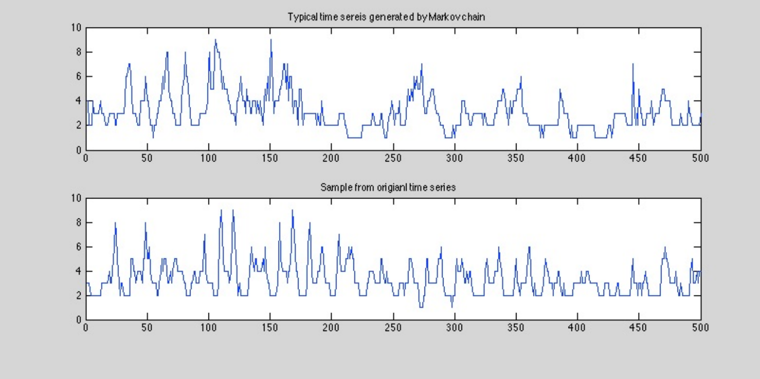
Empirical distribution from time series compared to the stationary distribution of chain:



$\boxed{\text{Step 10:}}$

Build a simulation of the Markov chain, using the transition matrix that you computed in Step 6. Generate a typical time series using your simulation, and compare with the original time series. Does it look like a good model?

Compare simulation of the Markov chain with original time series:



$\boxed{\text{Step 11:}}$ Compare your simulation with the original time series using the autocorrelation function. Given a time series $X_1, X_2, \ldots, X_N$, let $\overline{X}$ be the average, then for $k = 0, 1, 2, \ldots$ define

$$R(k) = \frac{\sum_{i=1}^{N-k} \left(X_i - \overline{X}\right) \left(X_{i+k} - \overline{X}\right)}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}$$

Compute the autocorrelation $R(0), R(1), R(2), \ldots$ for the original time series and for the time series you generated in Step 10. Does it look like they describe the same series?

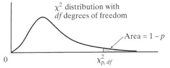
$\boxed{\text{Step 12:}}$ Compare your simulation with the original time series using a goodness of fit test for the 2-step transition probabilities. Namely, let $\widehat{p_{ij}}$ be the transition matrix that you computed in Step 6 above. Do the same kind of calculation on the original time series to get the frequency of going from each state i to each state j in *two steps*. Call this frequency N_{ij}, and let $N_i = \sum_j N_{ij}$. Compare these frequencies with the 2-step frequencies computed using $\widehat{p_{ij}}$, that is

$$M_{ij} = N_i\, q_{ij} = N_i \sum_k \widehat{p_{ik}}\, \widehat{p_{kj}}$$

Use a goodness of fit test to compare these at the 5% significance level for each state i (see notes on 'Goodness of Fit Test') and decide if the Markov chain $\{q_{ij}\}$ is a good model for the 2-step transitions of the data set.

$\boxed{\text{Step 13:}}$ write a report (maximum 7 pages) explaining how you carried out the above steps, including: source and nature of raw data set, how it was cleaned, choice of states for Markov chain, empirical distribution, transition matrix, stationary distribution, comparison of empirical and stationary distributions, autocorrelation function, goodness of fit test. At the end of your report, answer this question: ‘do you consider that the Markov chain method produces a good model for this time series? explain your answer’.

Table A.3 Upper and Lower Percentiles of χ^2 Distributions



	p							
df	0.010	0.025	0.050	0.10	0.90	0.95	0.975	0.99
1	0.000157	0.000982	0.00393	0.0158	2.706	3.841	5.024	6.635
2	0.0201	0.0506	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.336	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.688	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
31	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191
32	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486
33	17.073	19.047	20.867	23.110	43.745	47.400	50.725	54.776
34	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061