Ministério da
Ciência e Tecnologia

INPE-16608-RPQ/832

PROCESSO DE MINERAÇÃO DE DADOS NO ESTUDO DE FENÔMENOS SOLARES E GEOMAGNÉTICOS

Keila Silveira Corrêa

Relatório final da disciplina Princípios e Aplicações de Mineração de Dados (CAP-359) do Programa de Pós-Graduação em Computação Aplicada, ministrada pelo professor Rafael Santos.

Registro do documento original:
<http://urlib.net/sid.inpe.br/mtc-m18@80/2009/10.22.00.02>

INPE
São José dos Campos
2009

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3945-6911/6923
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

RESUMO

Este relatório descreve as etapas de um processo de mineração de dados aplicado na análise de fenômenos solares e geomagnéticos como a ocorrências de manchas solares, explosões solares, alterações no índice de atividade solar e perturbações geomagnéticas, procurando extrair padrões de comportamento ou estabelecer relações entre esses eventos.

LISTA DE FIGURAS

SUMÁRIO

Pág.

1 INTRODUÇÃO

O estudo do Clima Espacial refere-se ao estudo de fenômenos solares, ocorrências físicas no ambiente espacial e sua influência e impactos nos sistemas tecnológicos espaciais e terrestres. Dentre esses fenômenos, destacamos neste trabalho a análise das ocorrências de manchas e flares (explosões) solares, o grau de atividade solar e as perturbações geomagnéticas ocorridas no período entre 2001 e 2004.

Uma mancha solar é uma região onde ocorre uma redução de temperatura e pressão das massas gasosas no Sol. Possui intensos campos magnéticos e quanto maior suas quantidades, maiores são as alterações na ionosfera terrestre e, consequentemente sua influência nas comunicações de rádio e condições climáticas do planeta, entre outros efeitos.

Flares solares são explosões que ocorrem na superfície do Sol e que acontecem quando uma gigantesca quantidade de energia armazenada em campos magnéticos próximos às manchas solares é repentinamente liberada. Em períodos de maior intensidade solar, o número de manchas solares aumenta e consequentemente as explosões.

A intensidade solar é medida pelo índice de fluxo solar, também conhecido como índice F10.7, que correspondente às ondas de rádio emitidas pelo Sol cujo comprimento de onda é de 10,7 cm na freqüência de 2,8 GHz.

Os índices geomagnéticos Kp (Planetarische Kennziffer) e Dst (Disturbance Storm Time) medem o grau de perturbação do campo geomagnético terrestre. Os distúrbios na intensidade desse campo geomagnético causados por eventos solares são chamados tempestades magnéticas.

Este relatório descreve as etapas do processo de mineração de dados aplicadas na análise dos fenômenos solares e geomagnéticos descritos acima.

2 MATERIAIS E MÉTODOS

Para a realização deste trabalho foram utilizadas as seguintes ferramentas e softwares:
- sistema de gerenciamento de banco de dados PostgreSQL 8.3
- ambiente de desenvolvimento Java NetBeans IDE 6.7.1
- planilha Excel 2003
- programa Weka versão 3.7

Obedecendo as etapas de um processo de mineração de dados, as fases do trabalho foram:

- Obtenção dos dados
- Análise preliminar e seleção
- Pré-processamento e transformação dos dados
- Aplicação de algoritmos de mineração de dados
- Análise dos resultados

2.1 Obtenção dos dados

As informações escolhidas relacionadas aos fenômenos solares foram:
- número de manchas solares (SSN indice),
- explosões solares (solar flares)
- índice de atividade solar (F10.7 indice)

As informações escolhidas relacionadas ao comportamento geomagnético foram:
- índice Kp (Planetarische Kennziffer)
- índice Dst (Disturbance Storm Time)

Para o estudo foram escolhidos os dados medidos entre os anos de 2001 e 2004, período considerado de alta atividade solar.

Os índices solares estão disponíveis para *download* no site do NOAA's NGDC (National Geophysical Data Center) nos EUA no endereço ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA. Os índices geomagnéticos foram obtidos no site do World Data Centre for Geomagnetism (WDC-C2), Kyoto, Japão em http://swdcwww.kugi.kyoto-u.ac.jp/.

As figuras abaixo mostram exemplos do formato dos arquivos obtidos.

DAILY SUNSPOT NUMBERS 2001

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Yr Day
89	78	52	186	107	58	74	62	103	168	96	133	2001 01
94	78	53	166	118	99	83	81	106	144	100	149	2001 02
88	92	75	169	115	99	80	93	120	135	100	150	2001 03
98	91	92	134	132	96	71	115	108	132	111	145	2001 04
110	105	104	133	118	106	62	130	120	114	130	158	2001 05
130	110	91	110	92	119	45	120	141	104	140	142	2001 06
131	111	85	100	79	129	47	118	166	103	123	138	2001 07
105	111	63	115	55	142	54	117	182	77	152	140	2001 08
115	114	79	110	63	168	71	104	166	79	149	141	2001 09
101	105	97	114	60	159	70	99	150	98	149	115	2001 10
115	100	90	115	80	173	69	112	126	113	145	106	2001 11
117	71	95	103	84	171	90	112	149	127	121	117	2001 12
111	71	74	98	85	160	111	91	150	108	118	119	2001 13
92	68	80	92	102	180	99	93	148	115	118	101	2001 14
100	75	75	75	96	186	102	106	130	123	117	108	2001 15
75	73	70	58	99	191	113	127	121	121	90	120	2001 16
59	71	51	28	95	178	123	117	112	126	85	119	2001 17
60	76	61	38	93	153	127	106	136	131	92	115	2001 18
73	75	66	62	85	141	122	100	143	143	81	99	2001 19
61	76	80	86	82	136	118	101	183	160	87	101	2001 20
81	94	88	116	95	144	96	110	173	154	80	120	2001 21
93	81	85	109	121	151	100	112	164	135	87	135	2001 22
112	59	113	106	134	155	101	119	186	143	80	133	2001 23
118	56	149	109	118	145	90	116	200	135	67	157	2001 24
106	56	186	119	112	131	79	92	193	151	73	143	2001 25
84	58	218	119	118	114	61	101	175	154	84	167	2001 26
97	50	241	128	124	107	60	112	176	143	76	164	2001 27
102	51	235	107	103	89	63	121	170	139	107	156	2001 28
90		233	113	85	74	46	96	159	120	115	137	2001 29
70		231	112	75	65	57	99	165	103	121	134	2001 30
86		205		69		52	115		93		135	2001 31
95.6	80.6	113.5	107.7	96.6	134.0	81.8	106.4	150.7	125.5	106.5	132.2	

Values are final.

Figura 2.1 – Exemplo de formato de arquivos de dados de SSN

SOUTHERN HEMISPHERE

2001

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	1.42	0.00	0.32	1.48	0.00	0.40	0.00	2.41	0.15	1.14	2.26	24.00
2	0.00	0.00	0.00	0.23	0.00	1.04	0.00	1.48	8.34	2.76	1.38	1.96
3	0.55	0.00	0.09	5.98	0.00	3.56	0.00	2.46	0.35	0.98	0.34	5.51
4	0.09	0.55	0.13	3.41	0.00	2.96	0.00	1.20	4.77	0.77	2.28	7.11
5	0.23	0.08	0.68	13.30	0.00	4.65	0.00	7.94	6.57	3.08	2.94	3.74
6	1.59	0.00	0.00	2.76	0.00	3.24	0.02	2.36	8.62	3.17	4.39	11.66
7	0.00	0.20	1.19	0.63	0.00	0.58	0.47	7.28	1.92	1.40	10.75	7.42
8	0.13	0.00	0.00	0.17	0.00	8.81	2.01	1.55	8.34	0.00	16.58	2.15
9	0.00	3.14	5.31	13.67	0.00	17.33	2.80	2.67	33.10	12.51	23.18	5.45
10	0.00	1.94	2.93	29.93	0.56	4.67	0.25	0.50	6.49	1.66	10.54	6.84
11	0.00	0.68	0.71	14.59	0.37	1.80	0.09	0.84	4.58	0.25	6.69	3.60
12	0.74	0.99	1.94	19.01	12.60	1.60	0.06	0.27	4.89	1.27	1.07	10.26
13	0.00	0.00	0.18	0.13	3.90	4.84	0.13	0.28	7.43	0.73	2.70	10.32
14	0.00	0.00	1.07	3.04	1.66	0.00	3.88	0.77	5.45	0.38	4.82	4.10
15	0.00	0.00	1.68	13.19	3.19	4.12	6.41	0.11	13.73	0.00	1.70	0.49
16	0.00	0.00	1.47	0.09	1.76	0.77	7.85	0.13	22.12	0.17	0.60	0.96
17	0.00	0.00	3.66	0.00	2.40	0.00	1.08	0.26	10.35	1.02	34.81	0.10
18	0.00	0.00	4.29	0.00	2.32	0.00	0.78	0.29	1.79	4.36	0.72	0.42
19	2.03	0.28	0.58	0.00	0.41	1.32	8.14	0.04	0.82	7.51	2.99	1.46
20	26.49	0.56	6.54	0.00	0.00	0.00	0.72	0.91	0.88	2.28	4.52	0.11
21	1.57	1.06	3.81	0.44	0.14	0.00	1.65	0.00	1.46	2.21	8.21	1.93
22	0.82	1.01	0.00	0.00	0.57	1.22	1.90	3.50	5.69	28.32	24.25	1.99
23	0.37	1.69	1.60	1.10	0.02	0.23	1.38	2.61	4.57	6.22	0.87	6.95
24	1.49	0.00	2.75	0.09	0.44	0.00	0.84	10.53	27.15	1.22	12.03	11.41
25	1.76	0.09	0.50	0.00	2.24	0.63	0.23	46.25	12.42	47.43	4.08	5.88
26	2.24	0.10	3.66	0.00	0.00	0.00	0.00	4.40	5.55	2.39	0.33	8.44
27	0.49	0.00	10.20	0.08	0.00	3.00	1.39	6.80	3.31	1.44	9.18	35.09
28	10.91	0.26	2.34	0.33	0.00	1.59	2.95	3.59	21.81	0.00	3.35	3.98
29	1.79		0.38	0.00	0.00	0.41	0.38	2.48	0.62	0.45	1.12	1.57
30	0.34		3.81	0.00	0.00	0.05	0.00	9.84	2.33	0.00	3.22	1.05
31	0.02		1.73		0.24		1.46	0.71		0.00		3.34

Figura 2.2 – Exemplo de formato de arquivos de dados de *flare* solar

2001 OBSERVED DAILY SOLAR FLUX 2001
Penticton 2800 MHz Series C (Multiplied by Ten) 2000 UT

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	1710	1609	1314	2575	1845#	1330	1354	1202	1841	2165	2356	2213
2	1761	1663	1297	2280	1762	1340	1343	1208	1825	2009	2135	2450
3	1699	1636	1396	2231	1723	1453	1319	1316	1987	1917	2160	2350
4	1746	1481	1410	2048	1756	1538	1270	1484	2184	1865	2273	2333
5	1763	1653	1558	2075*	1606	1534	1196	1560	2183	1769	2346	2370
6	1794	1700	1578	1917#	1550	1577	1164	1637	2222	1804	2374	2467
7	1767	1640	1766	1795	1383	1648	1178	1663	2261	1727	2688	2259
8	1671	1565	1672	1692	1287	1802	1263	1669	2495	1712	2478	2205
9	1663	1624	1614	1648	1294	1770	1300	1633	2362	1764	2708	2242
10	1628	1607	1601	1697	1304	1630	1300	1603	2445	1787	2459	2190
11	1659	1513	1578	1596	1366	1624	1319	1650	2497	1748	2340	2206
12	1783	1446	1576	1490	1381	1664	1339	1598	2351	1792	2273	2367
13	1843	1413	1473	1370	1389	1814	1333	1515	2397	1795	2316	2202
14	1763	1379	1422	1387	1382	1947	1408	1473	2366	1919	2172	2166#
15	1692	1351	1361	1342	1421	1969	1421	1467	2193	1929	2070	2178
16	1619	1296	1399	1234	1378	2076	1498	1426	2071	2072	2021	2091
17	1519	1298	1342	1261	1474	2046	1456	1449	1991	2174	1985	2055
18	1515	1320	1398	1318	1382	2213	1430	1561	2038	2287	1882	2118
19	1525	1370	1470	1445	1413	1954	1423	1575	1988	2476	1913	2082
20	1532	1455	1533	1804	1415	1985	1426	1561	2268	2447	1850	2211
21	1515	1436	1594	1911	1501	2003	1390	1602	2386	2241	1842	2343
22	1621	1458	1830	1925	1520	2036	1404	1615	2552	2327	1900#	2428
23	1671	1452	1800	1964	1587	2062	1432	1697	2585	2264	1773	2546
24	1725	1373	2187	1935	1703	1948	1325	1749	2793	2387	1730	2745
25	1686	1349	2168	1939	1619	1824	1333	1990	2751	2389	1700	2588
26	1656	1354	2637	1962	1474	1679	1234	1899	2826	2365	1748	2678
27	1668	1306	2734	1908	1469	1479	1214	1920	2695	2465	1904	2746
28	1676	1318	2735	1878	1430	1402	1155	1992	2845	2272	1985	2633#
29	1654		2617	1917	1385	1399	1169	1970	2395	2158	2164	2644
30	1596		2568	1878	1323	1366	1145	1992	2358	2260	2258	2466
31	1533		2456		1328		1168	1887		2211		2456

Figura 2.3 – Exemplo de formato de arquivos de dados de índice F10.7

Os valores do índice Kp são informados a cada 3 h, enquanto que os valores do índice Dst são informados a cada hora (Figuras 2.4 e 2.5).

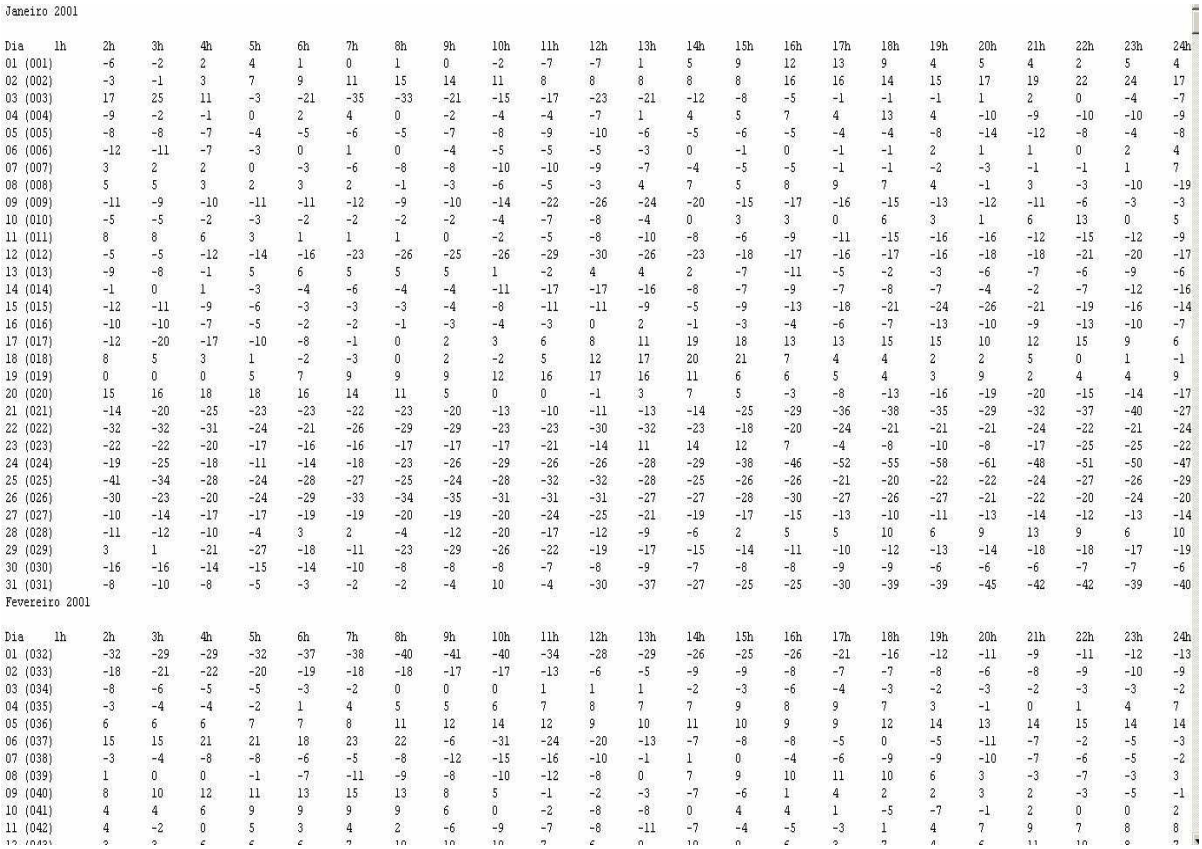
Janeiro 2001

Dia	0h	3h	6h	9h	12h	15h	18h	21h	Total	
01	(001)	0,0	0,3	1,0	1,0	0,3	0,3	0,7	0,7	4,3
02	(002)	1,3	0,3	0,0	0,0	0,7	0,3	1,0	2,7	6,3
03	(003)	2,3	3,7	2,3	2,3	2,3	0,7	0,7	1,0	15,3
04	(004)	2,3	1,3	3,0	2,3	3,0	3,0	2,3	2,3	19,7
05	(005)	1,7	0,7	0,7	0,3	0,7	0,3	1,3	3,0	8,7
06	(006)	2,0	1,0	1,0	0,7	1,0	1,0	0,3	1,0	8,0
07	(007)	2,0	1,7	1,3	0,7	0,3	1,0	1,0	2,0	10,0
08	(008)	2,0	2,0	1,0	1,7	1,0	2,0	4,0	3,7	17,3
09	(009)	2,0	0,7	1,0	1,7	1,7	1,0	1,3	1,3	10,7
10	(010)	0,3	0,3	0,3	0,3	1,0	3,0	2,0	2,3	9,7
11	(011)	1,7	0,3	0,7	1,0	1,7	2,7	2,7	2,7	13,3
12	(012)	3,3	3,0	2,7	1,7	1,0	0,3	0,0	0,3	12,3
13	(013)	1,0	1,0	1,7	2,3	2,7	1,3	0,7	1,3	12,0
14	(014)	2,3	2,7	2,3	1,7	1,0	1,0	2,3	2,0	15,3
15	(015)	0,7	0,0	1,3	2,0	2,3	1,3	1,7	3,0	12,3
16	(016)	1,0	0,3	1,3	1,3	3,0	1,7	1,0	1,0	10,7
17	(017)	3,0	1,7	0,7	0,3	1,3	2,3	2,0	1,3	12,7
18	(018)	0,0	0,0	1,3	2,0	1,3	0,7	1,0	2,0	8,3
19	(019)	1,7	1,3	0,3	0,0	1,0	1,0	1,3	2,0	8,7
20	(020)	0,7	0,7	2,7	2,3	2,3	2,3	2,7	3,3	17,0
21	(021)	2,3	2,7	2,7	4,0	3,3	3,7	3,7	4,0	26,3
22	(022)	3,7	3,3	2,3	2,3	2,3	2,7	2,0	2,7	21,3
23	(023)	2,0	1,7	0,3	4,0	3,0	3,3	4,7	4,0	23,0
24	(024)	2,7	3,0	1,0	3,3	4,0	4,0	4,7	3,0	25,7
25	(025)	1,7	1,3	0,7	0,7	0,3	1,7	2,7	2,0	11,0
26	(026)	2,0	2,3	3,7	2,7	2,0	2,0	2,3	2,7	19,7
27	(027)	2,3	2,0	1,0	0,7	0,7	0,7	0,7	0,7	8,7
28	(028)	0,3	1,0	1,3	1,7	3,0	2,0	3,0	2,7	15,0
29	(029)	4,7	5,0	2,0	1,3	1,0	1,0	2,3	2,7	20,0
30	(030)	2,0	2,0	0,3	0,3	0,3	0,7	0,0	0,0	5,7
31	(031)	0,3	0,7	4,0	4,0	3,7	3,7	4,0	3,0	23,3

Fevereiro 2001

Dia	0h	3h	6h	9h	12h	15h	18h	21h	Total	
01	(032)	2,0	3,3	2,7	2,0	1,3	1,0	2,0	2,0	16,3
02	(033)	3,0	2,0	1,0	1,0	0,3	1,0	1,0	1,3	10,7
03	(034)	0,3	0,0	0,3	0,3	0,7	0,3	0,0	0,0	2,0
04	(035)	0,0	0,0	0,0	0,3	0,7	0,3	1,0	0,0	2,3
05	(036)	0,0	0,0	0,0	0,7	0,7	0,7	1,3	2,0	5,3
06	(037)	3,0	2,3	4,3	2,7	1,3	2,7	3,0	2,7	22,0

Figura 2.4 – Exemplo de formato de arquivos de dados de índice Kp



Janeiro 2001

Dia	1h	2h	3h	4h	5h	6h	7h	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h	19h	20h	21h	22h	23h	24h
01 (001)		-6	-2	2	4	1	0	1	0	-2	-7	-7	1	5	9	12	13	9	4	5	4	2	5	4
02 (002)		-3	-1	3	7	9	11	15	14	11	8	8	8	8	8	16	16	14	15	17	19	22	24	17
03 (003)		17	25	11	-3	-21	-35	-33	-21	-15	-17	-23	-21	-12	-8	-5	-1	-1	-1	1	2	0	-4	-7
04 (004)		-9	-2	-1	0	2	4	0	-2	-4	-4	-7	1	4	5	7	4	13	4	-10	-9	-10	-10	-9
05 (005)		-8	-8	-7	-4	-5	-6	-5	-7	-8	-9	-10	-6	-5	-6	-5	-4	-4	-8	-14	-12	-8	-4	-8
06 (006)		-12	-11	-7	-3	0	1	0	-4	-5	-5	-5	-3	0	-1	0	-1	-1	2	1	1	0	2	4
07 (007)		3	2	2	0	-3	-6	-8	-8	-10	-10	-9	-7	-4	-5	-5	-1	-1	-2	-3	-1	-1	1	7
08 (008)		5	5	3	2	3	2	-1	-3	-6	-5	-3	4	7	5	8	9	7	4	-1	3	-3	-10	-19
09 (009)		-11	-9	-10	-11	-11	-12	-9	-10	-14	-22	-26	-24	-20	-15	-17	-16	-15	-13	-12	-11	-6	-3	-3
10 (010)		-5	-5	-2	-3	-2	-2	-2	-2	-4	-7	-8	-4	0	3	3	0	6	3	1	6	13	0	5
11 (011)		8	8	6	3	1	1	1	0	-2	-5	-8	-10	-8	-6	-9	-11	-15	-16	-16	-12	-15	-12	-9
12 (012)		-5	-5	-12	-14	-16	-23	-26	-25	-26	-29	-30	-26	-23	-18	-17	-16	-17	-16	-18	-18	-21	-20	-17
13 (013)		-9	-8	-1	5	6	5	5	5	1	-2	4	4	2	-7	-11	-5	-2	-3	-6	-7	-6	-9	-6
14 (014)		-1	0	1	-3	-4	-6	-4	-4	-11	-17	-17	-16	-8	-7	-9	-7	-8	-7	-4	-2	-7	-12	-16
15 (015)		-12	-11	-9	-6	-3	-3	-3	-4	-8	-11	-11	-9	-5	-9	-13	-18	-21	-24	-26	-21	-19	-16	-14
16 (016)		-10	-10	-7	-5	-2	-2	-1	-3	-4	-3	0	2	-1	-3	-4	-6	-7	-13	-10	-9	-13	-10	-7
17 (017)		-12	-20	-17	-10	-8	-1	0	2	3	6	8	11	19	18	13	13	15	15	10	12	15	9	6
18 (018)		8	5	3	1	-2	-3	0	2	-2	5	12	17	20	21	7	4	4	2	2	5	0	1	-1
19 (019)		0	0	0	5	7	9	9	9	12	16	17	16	11	6	6	5	4	3	9	2	4	4	9
20 (020)		15	16	18	18	16	14	11	5	0	0	-1	3	7	5	-3	-8	-13	-16	-19	-20	-15	-14	-17
21 (021)		-14	-20	-25	-23	-23	-22	-23	-20	-13	-10	-11	-13	-14	-25	-29	-36	-38	-35	-29	-32	-37	-40	-27
22 (022)		-32	-32	-31	-24	-21	-26	-29	-29	-23	-23	-30	-32	-23	-18	-20	-24	-21	-21	-21	-24	-22	-21	-24
23 (023)		-22	-22	-20	-17	-16	-16	-17	-17	-17	-21	-14	11	14	12	7	-4	-8	-10	-8	-17	-25	-25	-22
24 (024)		-19	-25	-18	-11	-14	-18	-23	-26	-29	-26	-26	-28	-29	-38	-46	-52	-55	-58	-61	-48	-51	-50	-47
25 (025)		-41	-34	-28	-24	-28	-27	-25	-24	-28	-32	-32	-28	-25	-26	-26	-21	-20	-22	-22	-24	-27	-26	-29
26 (026)		-30	-23	-20	-24	-29	-33	-34	-35	-31	-31	-31	-27	-27	-28	-30	-27	-26	-27	-21	-22	-20	-24	-20
27 (027)		-10	-14	-17	-17	-19	-19	-20	-19	-20	-24	-25	-21	-19	-17	-15	-13	-10	-11	-13	-14	-12	-13	-14
28 (028)		-11	-12	-10	-4	3	2	-4	-12	-20	-17	-12	-9	-6	2	5	5	10	6	9	13	9	6	10
29 (029)		3	1	-21	-27	-18	-11	-23	-29	-26	-22	-19	-17	-15	-14	-11	-10	-12	-13	-14	-18	-18	-17	-19
30 (030)		-16	-16	-14	-15	-14	-10	-8	-8	-8	-7	-8	-9	-7	-8	-8	-9	-9	-6	-6	-6	-7	-7	-6
31 (031)		-8	-10	-8	-5	-3	-2	-2	-4	10	-4	-30	-37	-27	-25	-25	-30	-39	-39	-45	-42	-42	-39	-40

Fevereiro 2001

Dia	1h	2h	3h	4h	5h	6h	7h	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h	19h	20h	21h	22h	23h	24h
01 (032)		-32	-29	-29	-32	-37	-38	-40	-41	-40	-34	-28	-29	-26	-25	-26	-21	-16	-12	-11	-9	-11	-12	-13
02 (033)		-18	-21	-22	-20	-19	-18	-18	-17	-17	-13	-6	-5	-9	-9	-8	-7	-7	-8	-6	-8	-9	-10	-9
03 (034)		-8	-6	-5	-5	-3	-2	0	0	0	1	1	1	-2	-3	-6	-4	-3	-2	-3	-2	-3	-3	-2
04 (035)		-3	-4	-4	-2	1	4	5	5	6	7	8	7	7	9	8	9	7	3	-1	0	1	4	7
05 (036)		6	6	6	7	7	8	11	12	14	12	9	10	11	10	9	9	12	14	13	14	15	14	14
06 (037)		15	15	21	21	18	23	22	-6	-31	-24	-20	-13	-7	-8	-8	-5	0	-5	-11	-7	-2	-5	-3
07 (038)		-3	-4	-8	-8	-6	-5	-8	-12	-15	-16	-10	-1	1	0	-4	-6	-9	-9	-10	-7	-6	-5	-2
08 (039)		1	0	0	-1	-7	-11	-9	-8	-10	-12	-8	0	7	9	10	11	10	6	3	-3	-7	-3	3
09 (040)		8	10	12	11	13	15	13	8	5	-1	-2	-3	-7	-6	1	4	2	2	3	2	-3	-5	-1
10 (041)		4	4	6	9	9	9	9	6	0	-2	-8	-8	0	4	4	1	-5	-7	-1	2	0	0	2
11 (042)		4	-2	0	5	3	4	2	-6	-9	-7	-8	-11	-7	-4	-5	-3	1	4	7	9	7	8	8

Figura 2.5 – Exemplo de formato de arquivos de dados de índice Dst

Foi implementado um aplicativo em linguagem Java para ler os diferentes formatos mostrados acima e armazenar as informações em uma base de dados criada no SGBD PostgreSQL.

A tabela *tabdatasolargeo* foi gerada com 2.103.840 tuplas contendo dados dos anos de 2001 a 2004, sendo a sua estrutura mostrada na Figura 2.6.

yy	mm	dd	hh	ssn	flares	f10	dst	kp

Figura 2.6 – Tupla da tabela tabdatasolargeo

Onde:

- yy, mm, dd, hh são ano, mês, dia e hora da coleta
- ssn – número de manchas solares (sunspot number)
- flares – *flares* solares

- f10 – valor do índice f10.7
- dst e Kp – valores de dst e Kp respectivamente.

2.2 Análise preliminar e seleção

A análise preliminar tem o objetivo de identificar a necessidade de tratamentos ou outros processamentos nos dados existentes e selecionar períodos de amostras de dados que apresentem comportamentos relevantes relacionados à ocorrência de um fenômeno especifico. No caso deste trabalho, foi escolhido o fenômeno das tempestades magnéticas, representado pelo índice Dst.

Tabela 2.1 – Valores do índice Dst para diferentes intensidades de tempestade magnética

Muito intensa	<-250
Intensa	-100 a -250
Moderada	-50 a -100
Fraca	-30 a -50
Inexistente	>-30

Com a inserção dos dados no SGBD foi possível, por meio de consultas SQL, extrair informações e gerar gráficos mostrando a variabilidade dos valores na linha do tempo. Nesta fase utilizou-se a ferramenta Excel.

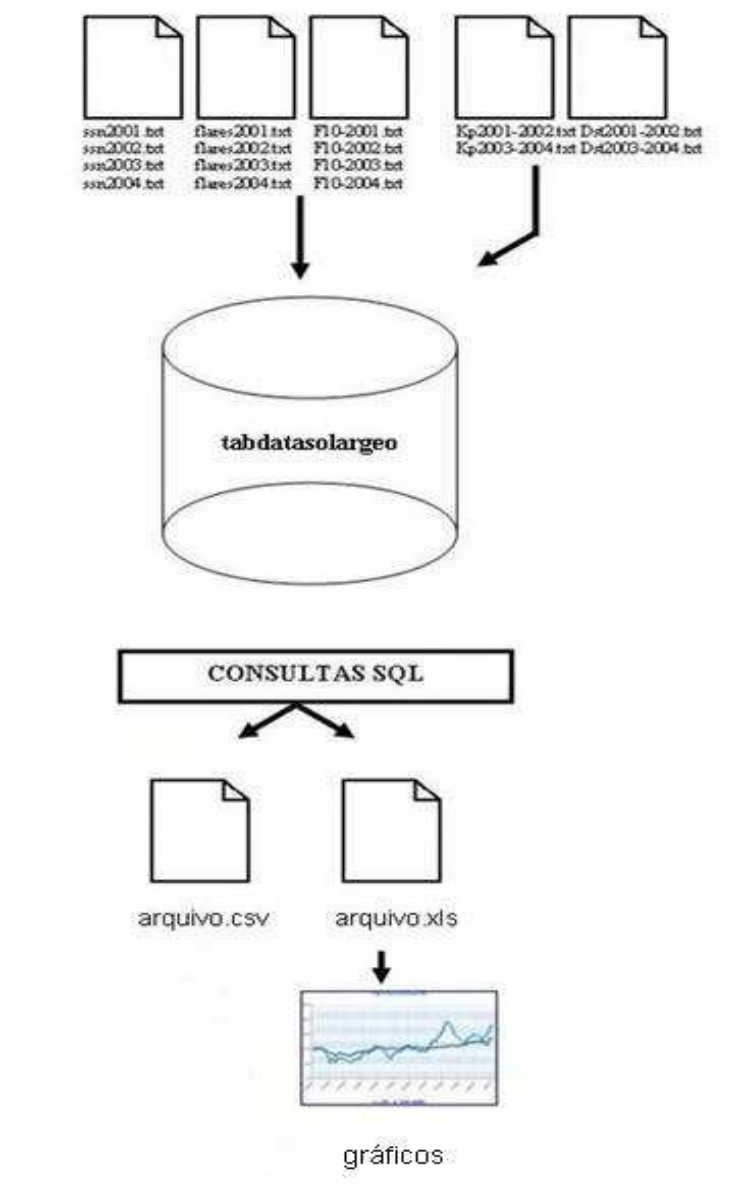


Figura 2.7 – Processo de obtenção de dados e analise preliminar

Análise 1 – período 2001 a 2004 – dados informados por dia
Para esta análise foram calculados os índices Kp e Dst médios para cada dia.

Sentença SQL: *SELECT yy, mm, dd, ssn, flares, f10, avg(dst) as "m-dst", avg(kp) as "m-kp" FROM tabdatasolargeo GROUP BY yy, mm, dd, ssn, flares, f10 ORDER BY yy,mm,dd;*

Saída: 1.461 registros

Arquivos gerados: solargeo-2001-2004-diario.xls e solargeo-2001-2004-diario.csv

Resultado: Percebeu-se a relação entre o aumento de manchas solares, *flares* e a intensidade solar medida por F10.7 (Figuras 2.8, 2.9 e 2.10).

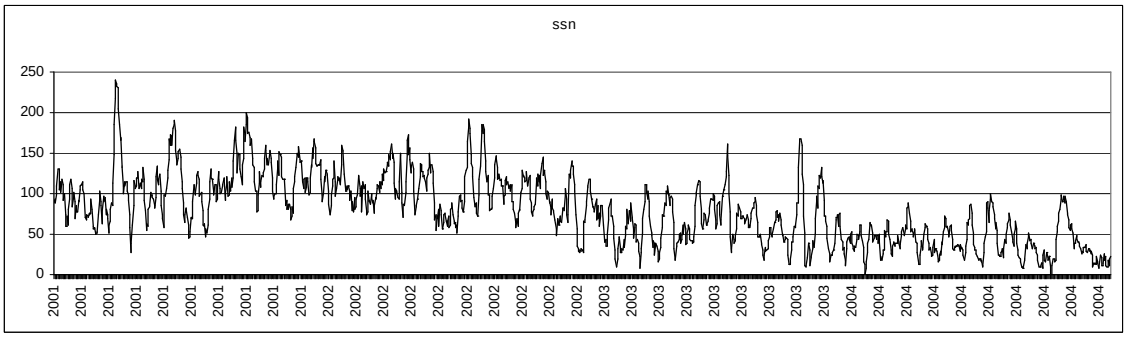


Figura 2.8 – Número de manchas solares no período de 2001 a 2004

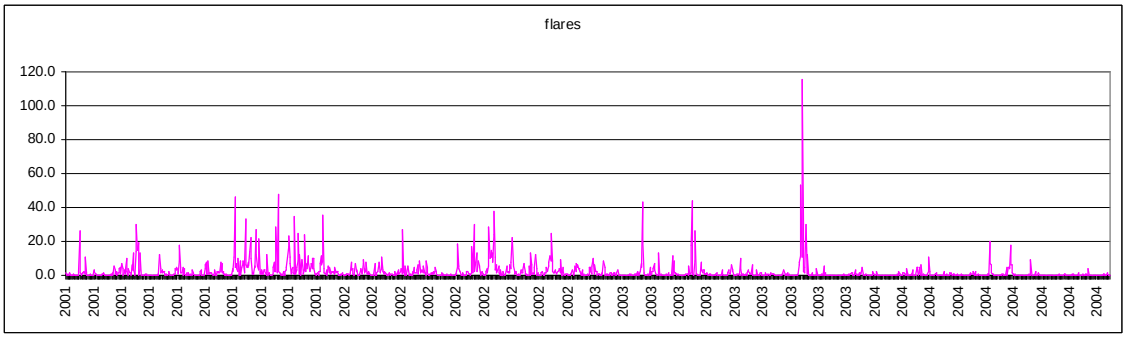


Figura 2.9 – Ocorrências de *flares* solares no período de 2001 a 2004

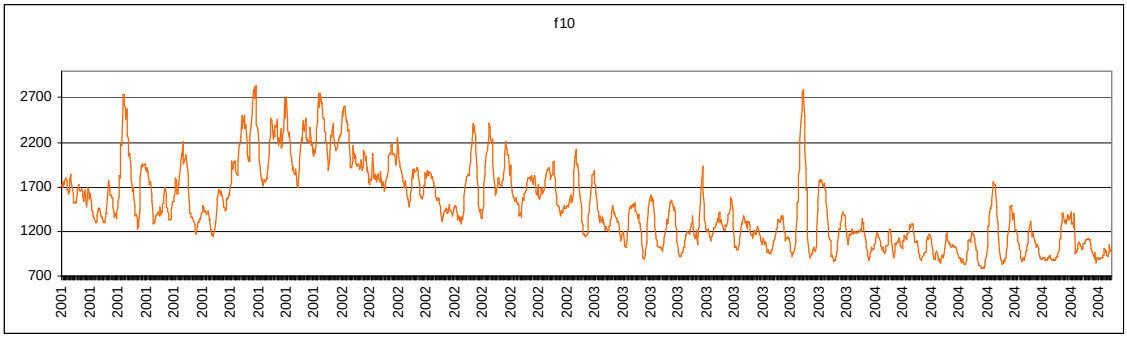


Figura 2.10 – Índice F10.7 no período de 2001 a 2004

O gráfico apresentado na Figura 2.11 mostra períodos de ocorrência de tempestades magnéticas, isto é, períodos onde o valor do atributo Dst esteve abaixo de -100.

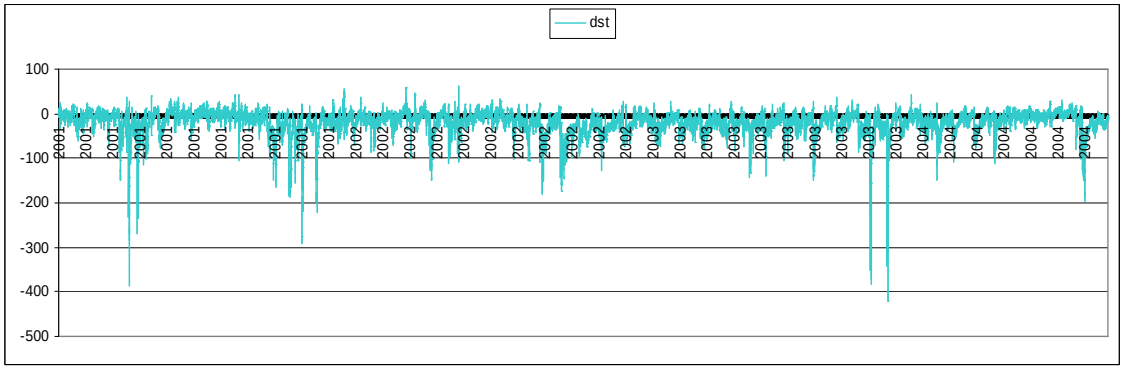


Figura 2.11 – Índice Dst no período de 2001 a 2004

Análise 2 – ano de 2001 – dados informados por hora

Sentença SQL: *SELECT yy, mm, dd, hh, ssn, flares, f10, dst, kp FROM tabdatasolargeo WHERE yy=2001 ORDER BY yy,mm,dd,hh;*

Saída: 525.600 registros

Arquivos gerados: solargeo-2001-hora.xls, solargeo-2001-hora.csv

Resultado: Percebeu-se mais claramente uma relação no comportamento dos atributos SSN (número de manchas solares) e f10 (índice F10.7), mostrada nas Figuras 2.12 e 2.13.

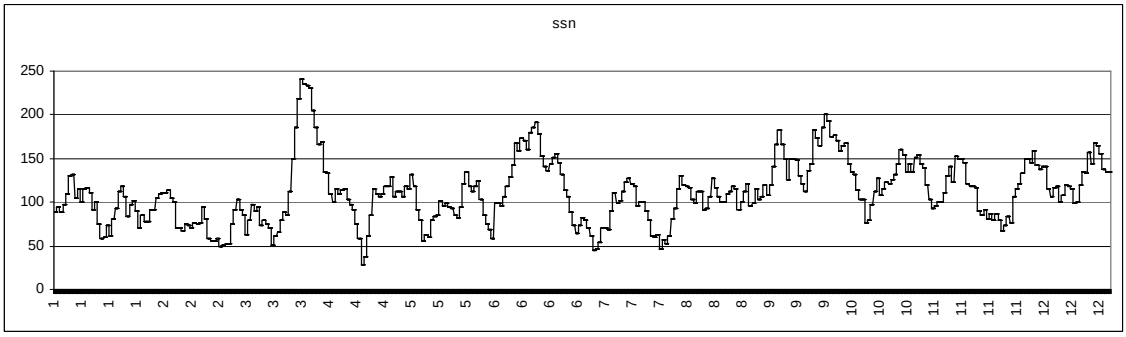


Figura 2.12 – Número de manchas solares – ano de 2001

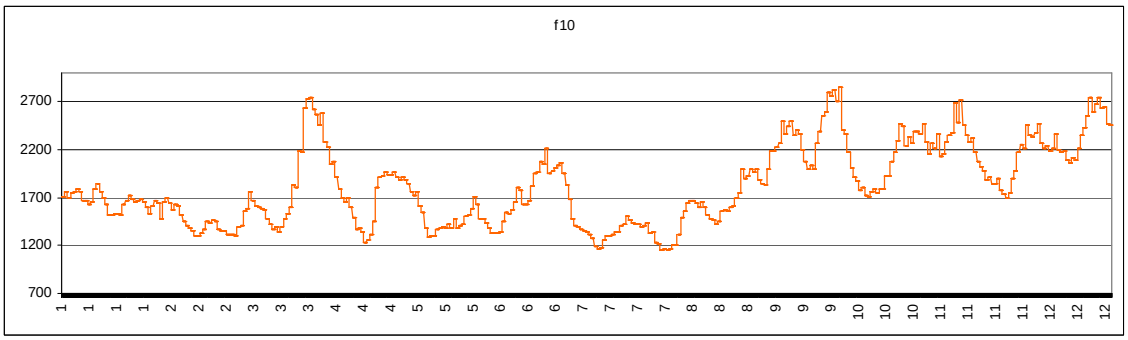


Figura 2.13 – Índice F10.7 – ano de 2001

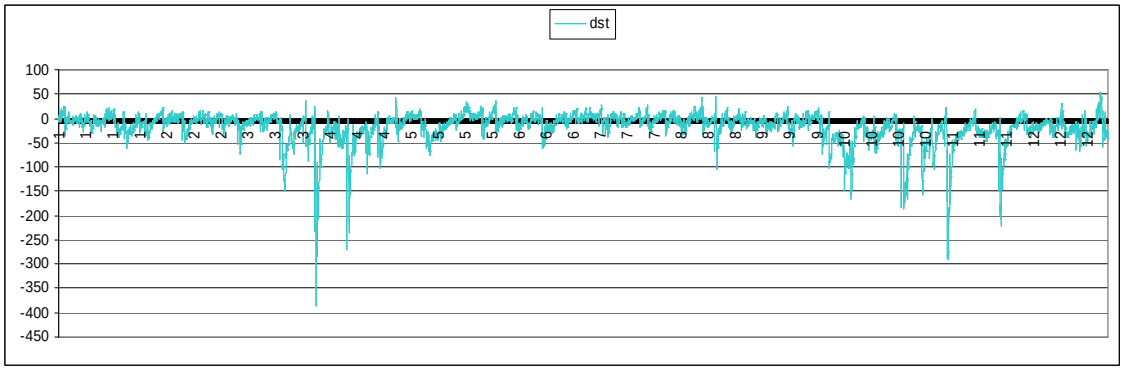


Figura 2.14 – Ocorrência de tempestades magnéticas no ano de 2001

Análise 3 – período de setembro a dezembro de 2001 – dados informados por hora

Sentença SQL: *SELECT yy, mm, dd, hh, ssn, flares, f10, dst, kp FROM tabdatasolargeo WHERE yy=2001 and mm>8 ORDER BY yy,mm,dd, hh;*

Saída: 175.680 registros

Arquivos gerados: solargeo-set-dez-2001-hora.xls e solargeo-set-dez-2001-hora.csv

Resultado: Observou-se uma relação inversa entre os índices Dst e Kp (Figura 2.15).

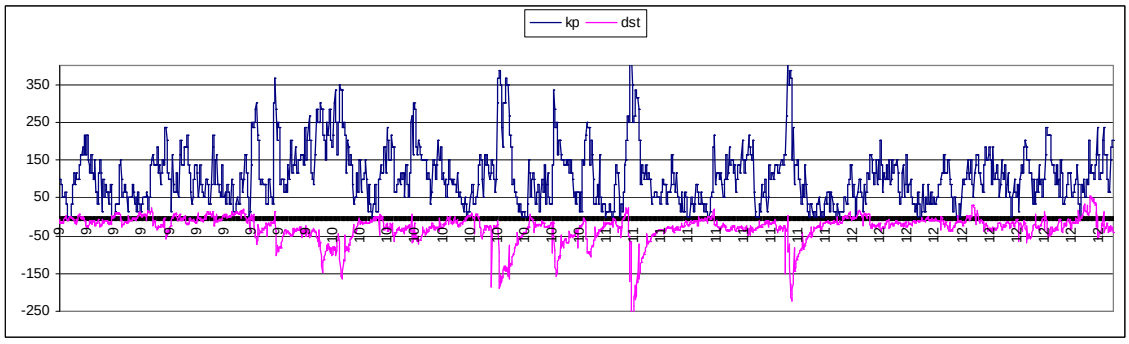


Figura 2.15 – Índices Kp e Dst – período de setembro a dezembro de 2001

2.3 Pré-processamento

O valor dos atributos SSN, flares e f10 foram suavizados por meio do cálculo do valor da média móvel de três dias.

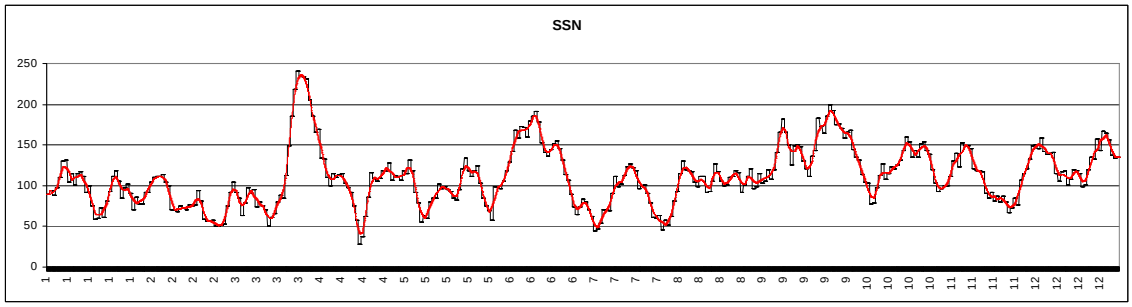


Figura 2.16 – Atributo ssn suavizado pela média móvel de 3 dias

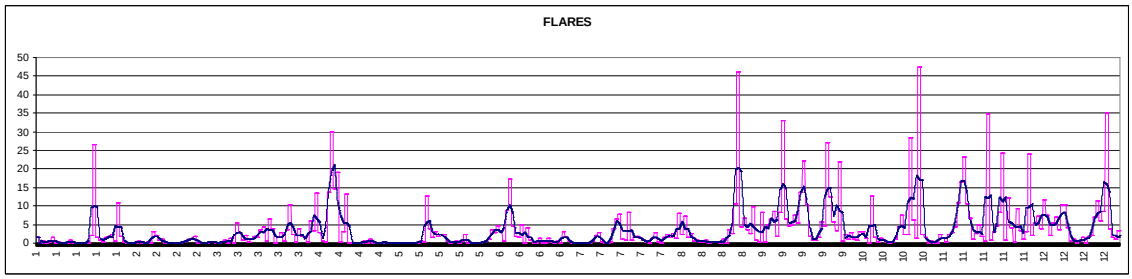


Figura 2.17 – Atributo “flares” suavizado pela média móvel de 3 dias

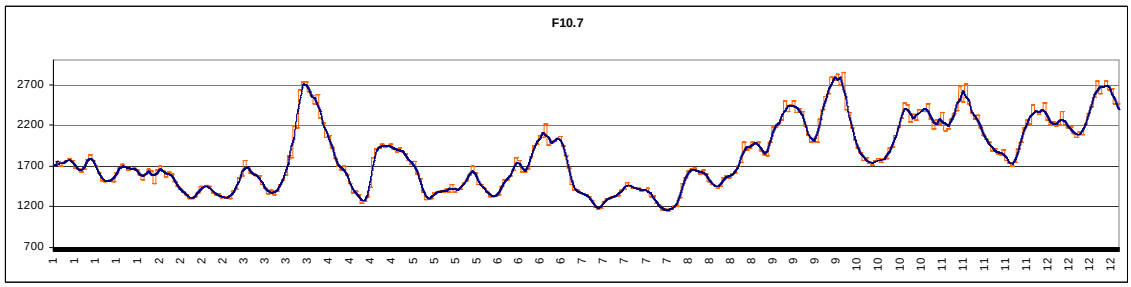


Figura 2.18 – Atributo f10 suavizado pela média móvel de 3 dias

Acrescentou-se ao processo os atributos *sazonalidade* (período do ano em que ocorreu o evento) e *horário*. Além disso, todos os atributos foram discretizados recebendo valores nominais, conforme a tabela abaixo.

Tabela 2.2 – Tabela de atributos e valores

atributo	**valores**
SSN	0-100, 100-200, >200
flares	0-40, 40-80, >80
f10	<110, 110-210, 210-250, >250
Kp	se <4, calmo, se >=4, perturbado
Dst	<-100, tempestade intensa -50 a -100 tempestade moderada -30 a -50, tempestade fraca

	>-30, nula
sazonalidade	saz1, ocorrência entre os meses de janeiro e março saz2, ocorrência entre abril e junho saz3, ocorrência entre julho e setembro saz4, ocorrência entre outubro de dezembro
horário	h1, entre 0h e 6h h2, entre 6h e 12h h3, entre 12h e 18h h4, entre 18h e 24h

Os arquivos .csv resultantes das consultas SQL (Figura 2.7) foram convertidos para o formato .arff e utilizados no software Weka na forma exemplificada na Figura 2.19 .

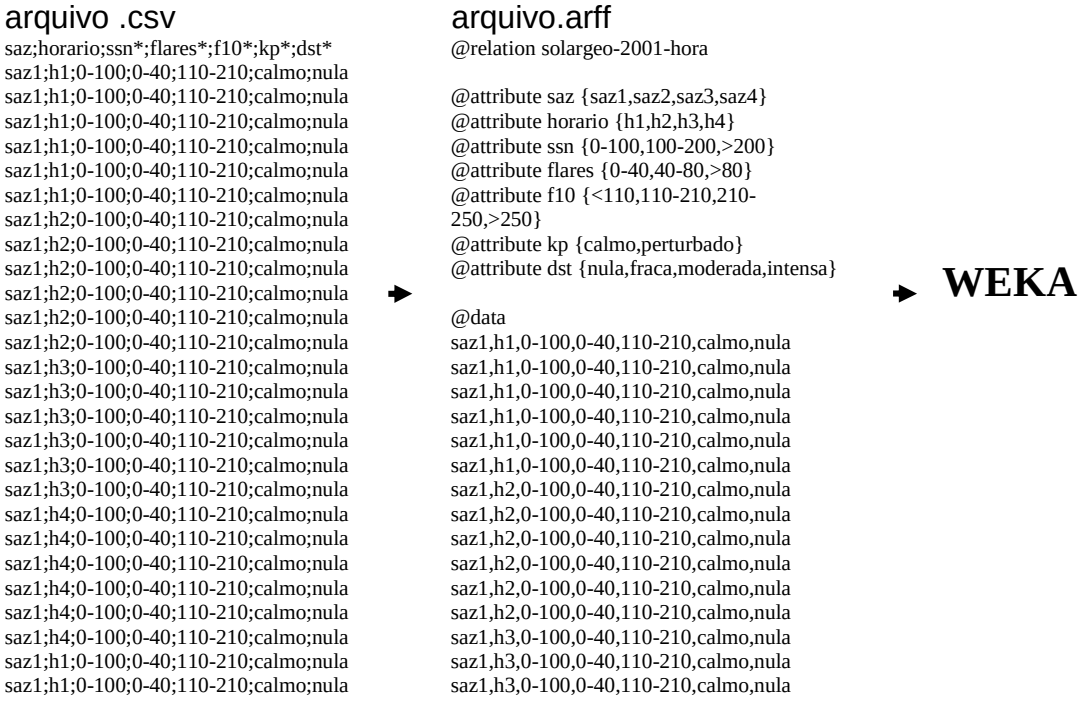


Figura 2.19 – Conversão de formato csv para arff

2.4 Aplicação de algoritmos de mineração de dados

Os algoritmos de mineração de dados utilizados foram o J48 e Apriori/PredictiveApriori do software Weka versão 3.7.

O algoritmo J48 é uma implementação do algoritmo de árvore de decisão C.45 (Quinlan, 1993). Trabalha gerando as regras com base na análise de cada atributo, através dos seus valores e da sua relação com os demais parâmetros evolvidos. Uma árvore de decisão é uma estrutura simples onde nós não terminais representam testes sobre um ou mais atributos e nós terminais refletem resultados de decisão. O algoritmo genérico para a criação de uma árvore de decisão primeiramente testa todos os atributos para a criação de um nó da árvore. É escolhido o atributo que tem maior ganho de informação. Divide-se a árvore em nós e sub-árvores usando o atributo selecionado. E é repetido, recursivamente, até não ser mais possível decidir por atributos.

O algoritmo Apriori, sendo um algoritmo de regras de associação, procura identificar relações e dependências significativas entre os atributos.
O algoritmo Predictive Apriori é derivado do algoritmo Apriori (SCHEFFER, T. et al., 2001). Combina as medidas de confiança e suporte numa única medida de confiança preditiva e encontra as melhores regras ordenadamente.

Os algoritmos foram aplicados aos seguintes conjuntos de dados:

a) período de 2001 a 2004 (Figura 2.20)
b) ano de 2001 (Figura 2.21)
c) meses de setembro a dezembro de 2001 (Figura 2.24) – período de ocorrências de tempestades magnéticas intensas

Algoritmo: weka.classifiers.trees.J48 -C 0.25 -M 2
Total Number of Instances 35064
Correctly Classified Instances 26809 76.4573 %
Incorrectly Classified Instances 8255 23.5427 %

=== Confusion Matrix ===

a	b	c	d	<-- classified as
25930	299	289	35	a = nula
4867	**305**	301	19	b = fraca
1772	135	**439**	51	c = moderada
295	23	169	**135**	d = intensa

Precision	Class
0.789	nula
0.4	fraca
0.366	moderada
0.563	intensa

Figura 2.20 – Resultado do algoritmo J48 aplicado ao conjunto de dados do período de 2001 a 2004

Devido ao grande número de instâncias no conjunto de dados anterior e também de eventos onde o valor de Dst = nula, optou-se por reduzir o período de representação dos dados para um ano (ano de 2001). Reduzindo o número de amostras para 12 meses, obteve-se maior precisão principalmente na classificação das tempestades moderadas e intensas (Figura 2.21)

Algoritmo: weka.classifiers.trees.J48 -C 0.25 -M 2
Total Number of Instances 8760
Correctly Classified Instances 7045 80.4224 %
Incorrectly Classified Instances 1715 19.5776 %

=== Confusion Matrix ===

a	b	c	d	<-- classified as
6787	60	29	20	a = nula
923	**102**	18	17	b = fraca
425	51	**54**	34	c = moderada
106	14	18	**102**	d = intensa

Precision	Class
0.824	nula
0.449	**fraca**
0.454	**moderada**
0.59	**intensa**

Figura 2.21 – Resultado do algoritmo J48 aplicado ao conjunto de dados do ano de 2001

Desprezando-se o atributo *horário* simplificou-se a árvore de decisão sem grande alteração nos resultados (Figura 2.22).

```
Algoritmo:weka.classifiers.trees.J48 -C 0.25 -M 2
(sem o atributo horario)
Total Number of Instances              8760
Correctly Classified Instances         7029     80.2397 %
Incorrectly Classified Instances       1731     19.7603 %

=== Confusion Matrix ===
   a    b    c    d   <-- classified as
 6786   63   28   19 |    a = nula
  934   91   20   15 |    b = fraca
  425   54   65   20 |    c = moderada
  114   14   25   87 |    d = intensa

               Precision   Class
                   0.822     nula
                    0.41    fraca
                   0.471     moderada
                   0.617    intensa
```

Figura 2.22 – Resultado do algoritmo J48 aplicado ao conjunto de dados do ano de 2001, desprezando-se o atributo *horário*

Aplicou-se o algoritmo J48 a dados do período de setembro a dezembro de 2001 (período de alta ocorrência de tempestades magnéticas) e os resultados foram melhores para a predição de tempestades intensas, precisão de 66% contra 59% do experimento anterior.

Melhores resultados foram obtidos mantendo-se os valores de SSN, flares e F10.7 numéricos e não discretizados (Figura 2.23).

Algoritmo:weka.classifiers.trees.J48 -C 0.25 -M 2
(atributos ssn, flares e F10.7 numéricos, **não** discretizados)

Total Number of Instances	8760	
Correctly Classified Instances	8018	91.5297 %
Incorrectly Classified Instances	742	8.4703 %

=== Confusion Matrix ===

a	b	c	d	<-- classified as
6677	187	26	6	a = nula
291	**702**	64	3	b = fraca
49	62	**440**	13	c = moderada
13	7	21	**199**	d = intensa

Precision	Class
0.95	nula
0.733	fraca
0.799	moderada
0.9	intensa

Figura 2.23 – Resultado do algoritmo J48 aplicado ao conjunto de dados do ano de 2001, com atributos numéricos (não discretizados)

3 CONCLUSÃO

O relatório descreveu as etapas de um exercício de mineração em um conjunto de dados que representava fenômenos dos meios solar e geomagnetico. Os processos de tratamento como limpeza, seleção, redução e pre-processamento de dados foram bem explorados e procurou-se obter os melhores resultados.
No entanto, a escolha dos atributos se baseou na facilidade e disponibilidade dos dados (dados estarem disponíveis para acesso na Internet) além da sua representatividade física. Para resultados mais satisfatórios, poderiam ter sido considerados outros atributos relacionados ao meio geomagnético como o campo magnétio interplanetário por exemplo, como também a aplicação de outros algoritmos de mineração, capazes de tratar a ocorrência de eventos não simultâneos.

REFERÊNCIAS BIBLIOGRÁFICAS

MURALIKRISHNA, A. Previsão do índice dst utilizando redes neurais artificiais e árvores de decisão. São José dos Campos. 132 p. Tese (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2009.

NOAA's NGDC - ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA - acessado em 1 de agosto de 2009.

REZENDE, L.F.C. Mineração de dados aplicada a análise e predição de cintilação ionosférica. São José dos Campos. 148 p. Tese (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2009.

QUINLAN, J. Ross. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

SCHEFFER, T. Finding Association Rules that Trade Support Optimally Against Confidence, in Principles of Data Mining and Knowledge Discovery. 5th European Conference, PKDD 2001, Freiburg, Alemanha, ,Proceedings, 3-5 Setembro, 2001.

Software Weka - http://www.cs.waikato.ac.nz/ml/weka

WDC-C2 - http://swdcwww.kugi.kyoto-u.ac.jp/ - acessado em 1 de agosto de 2009.