

小特集 核融合プラズマ研究におけるデータマイニングの活用

3. データマイニングを用いたプラズマ実験データの解析例

3. Examples of Data Mining Applications in Plasma Experiments

3.3 ネットワーク可視化を用いた新しい時系列解析法

3.3 A New Method for Time Series Analysis Using Network Visualization

谷澤 俊弘¹⁾, 中村 知道²⁾

TANIZAWA Toshihiro¹⁾ and NAKAMURA Tomomichi²⁾

¹⁾高知工業高等専門学校電気情報工学科, ²⁾兵庫県立大学大学院シミュレーション学研究科

(原稿受付: 2015年11月20日)

時系列データから線形モデルを生成し、その構造をネットワークによって可視化することにより、その時系列データが持つ潜在的な力学的性質を直観的に理解する方法を紹介する。基礎となる線形モデルは、通常使われる自己回帰モデル (auto-regressive model: AR model) を情報理論的に洗練した「縮約された自己回帰モデル」(reduced auto-regressive model: RAR model) である。この方法によって得られるネットワークのトポロジーには、時系列データが内包する力学的性質が反映される。この方法はどのような時系列データの解析にも用いることができ、核融合プラズマ乱流の解析にも適用可能である。

Keywords:

time series analysis, linear model, complex network, information criterion, visualization

3.3.1 はじめに

時系列データとはある物理量や現象の時間的な変化を一定の時間間隔で継続して観測して得られる数値の系列である。われわれの身の回りのさまざまな動的現象はすべて時系列データとして表される。

多数の要素から構成される複雑なシステムが生み出す動的現象を理解するためには、システムを特徴づけるいくつかの物理量相互の依存関係をできるだけ正確に特定し、それらの時間依存性を記述する方程式を導出してそのふるまいを再現し予測できなければならない。時間依存性を記述する方程式群の導出については、力学的運動方程式やマクスウェル方程式などの基本方程式から出発する微視的な導出とマクロな物理量の時間依存性を直接取り扱う現象論的導出に大別される。相互作用する多体系では、微視的記述は困難である場合が多く、そのような場合には現象論的記述が有効であり、そこから微視的モデル構築の手がかりも得られることがある。

本節では、現象論的モデル構築の一つの方法として、時系列データから統計モデルを直接作り、さらにそれをネットワークによって可視化する新たなアプローチを紹介する[1]。これにより、時系列データが持つ潜在的な力学的性質を直感的に理解できるようになる。時系列解析で通常用いられるモデルは自己回帰モデル (auto-regressive model: AR model) であるが、ここでは AR モデルよりも多くの利

点を持つ「縮約された自己回帰モデル」(reduced AR model: RAR model) を用いる。RAR モデルの利点や構築法、さらに、そのモデルをネットワーク化することで何がえてくるかについて焦点を当て、その方法のプラズマ乱流現象の解析への応用の可能性についても述べる。

3.3.2 時系列データと統計モデル

あるシステムの時刻 t での動的ふるまいが n 種類の物理量によって記述されているとする。これらの物理量はある一定の時間間隔 Δt で測定され、時刻 t は整数のインデックス k を用いて $t \equiv k\Delta t$ と表される。さらに、簡単のために時刻 $t+k\Delta t$ を単に $t+k$ と表すこととする。統計モデルとは、ある時刻の物理量 $x_i(t)$ をそれ以前の時刻の物理量を用い、ある統計誤差の範囲内で

$$x_i(t) = F(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n) \quad (1)$$

と表されるものである。ここで、 $\vec{v}_i = \{x_i(t-1), x_i(t-2), \dots\}$ は i 番目の物理量の時刻 t 以前の値であり、関数 F が $\{\vec{v}_i\}$ の線形関数であるか非線形関数であるかにしたがって、この統計モデルを線形モデル、非線形モデルと呼ぶ。本節では、議論を簡単な単一の物理量に関する線形モデルに限定し、自己回帰 (AR) モデル

corresponding author's address: Kochi National College of Technology Nankoku, KOCHI, 783-8508, Japan

corresponding author's e-mail: tanizawa@ee.kochi-ct.ac.jp

$$x(t) = a_0 + \sum_{l=1}^m a_l x(t-l) + \varepsilon_t \quad (2)$$

とその改良形である「縮約された自己回帰 (RAR) モデル」について詳述する。ここで、 ε_t はフィッティングの際の誤差項であり、ここではガウス型白色雑音となる。

AR モデルは長期的なゆらぎ (トレンド) を持たない定常的な時系列データの記述に広く用いられてきた[2]。モデリングに用いる十分に長い時系列データを $\{x(0), x(1), \dots, x(N-1), x(N)\}$ とすると、このデータに関して構築される AR モデルは $\hat{x}(t) = a_0 + a_1 x(t-1) + a_2 x(t-2) + \dots + a_m x(t-m)$ という形を取る。ここで、パラメータ $\{a_l\}$ は二乗誤差 $\sum_{t=0}^N |x(t) - \hat{x}(t)|^2$ が最小となるように決定される。この AR モデルの構築においては、最大の時間遅れ m をどう取るかが重要である。一般に m を大きく (小さく) するにつれて二乗誤差は小さく (大きく) なり、モデリングに用いた時系列データとの当てはまりは良く (悪く) なる。ここで、観測から得られるデータには、常に観測ノイズが含まれていることを考慮しなければならない。時系列データに完全に一致するようにモデルを構築すると、一般に m は非常に大きくなり、極端な場合、もとの時系列データ数と同じになるが、そのようなモデルは、その時系列データの持つ本質的な物理的性質だけではなく、そこに含まれる観測ノイズをも反映してしまう。モデル構築に際しては、本質的な現象のみを反映し、観測ノイズの影響は極力排除したい。このような良いバランスの取れる時間遅れ m の最適値を見つけていくことが重要となる。

最大時間遅れの決定には、情報量基準が用いられることが多い。最もよく知られているものは、赤池弘次氏によって提案された Akaike information criterion (AIC) [3] である。AIC が提案された後、様々な背景を持つ情報量基準が複数提案されている。例えば、Schwartz information criterion (SIC) [4]、Description Length, Predictive Description Length, Judd と Mees によって提案された Description Length, Normalized Maximum Likelihood などが挙げられる[5]。ここでは、AIC と Judd と Mees によって提案された Description Length (DL) [6] を紹介する。AIC は、本節のパラメータを用いると

$$\text{AIC}(m) = N \ln \frac{|\bar{e}|^2}{N} + 2m \quad (3)$$

となり、記述長 $\text{DL}(m)$ は

$$\text{DL}(m) = \left(\frac{N}{2} - 1 \right) \ln \frac{|\bar{e}|^2}{N} + (m+1) \left(\frac{1}{2} + \ln \gamma \right) - \sum_{j=1}^m \ln \delta_j \quad (4)$$

となる。ここで、 $\gamma = 32$ であり、 δ_j は、モデルの j 番目のパラメータ a_l を決める際に必要な精度を表す量である。また、 $\bar{e} = \{x(t) - \hat{x}(t)\}_{t=0}^N$ は近似誤差である。いずれの情報量基準においても、第 1 項の誤差を小さくするために m を大きくしていくと、やがて第 2 項の影響で情報量基準は増加

に転ずる。時間遅れの最大値 m はこの情報量基準を最小にする値として決定される。

3.3.3 RAR モデル

現在 AR モデルは有効な統計モデルとして広く用いられているが、最近、さらにそのモデル中の本質的な時間遅れ項のみを用いてモデルを構築する方法が、Judd と Mees らによって提案された[6, 7]。そのモデルは、 $x(t)$ を記述する際に、ある最大時間遅れ m までのすべての項 $\{x(t-1), x(t-2), \dots, x(t-m)\}$ を用いるのではなく、さらにその部分集合 $\{x(t-l_1), x(t-l_2), \dots, x(t-l_m)\}$ も考慮し、モデルに重要な役割を果たす項を情報量基準によって選び出すもので、縮約された自己回帰モデルと呼ばれる。

本節では図 1 に表されている西暦 1700 年から 2013 年までの一年ごとの太陽黒点数のデータを用いて、RAR モデルの構築法とその利点を具体的に述べる。

この時系列データが

$$x(t) = a_0 + a_{l_1} x(t-l_1) + a_{l_2} x(t-l_2) + \dots + a_{l_m} x(t-l_m) + \varepsilon_t \quad (5)$$

とモデル化されるとする。右辺には、ガウシアン雑音を別にすると、定数項と $\{l_1, l_2, \dots, l_m\}$ の時間遅れ項、合わせて $m+1$ 個の項がある。理想的には、可能なあらゆる $\{l_1, l_2, \dots, l_m\}$ の選び方の中から情報量基準を最小とするモデルを選び出せばよい。 m が 30 程度であれば、全探索を行ったとしてもさほど時間はかからない。しかし、それ以上の大きさになると非常に長い時間が必要となり、全探索は現実的ではなくなる。そのため、効率的に良いモデルを探索または構築する手法がいくつか提案されている[1, 6, 7]。この場合、得られるモデルは最適である保障はないため、準最適モデルとして扱われることとなる。ここでは、 $M = 20$ 、すなわち、ある時点での太陽黒点数のデータは最大 20 年前までのデータを用いて記述されると仮定し、全探索によりモデル構築を始める。したがって、モデル構築に用いる項の種類は、定数項と 20 までの時間遅れ項を合わせた 21 種類である。この 21 種類の項から、 $m = 1$ 個を用いたモデル (${}_{21}C_1 = 21$ 種類)、 $m = 2$ 個を用いたモデル (${}_{21}C_2 = 210$ 種類)、 $\dots$ 、 $m = 21$ 個すべてを用いたモデルについて、最小二乗誤差を用いたフィッティングによりパラメータを決定した上で情報量基準を計算し、その最小値を与えるものを、この時系列データの RAR モデルとして採用する。ここでは情報量基準として、式(4)の記述長 $\text{DL}(m)$ を用いた。この方法により図 1 のデータを記述する RAR モデルを構築すると、

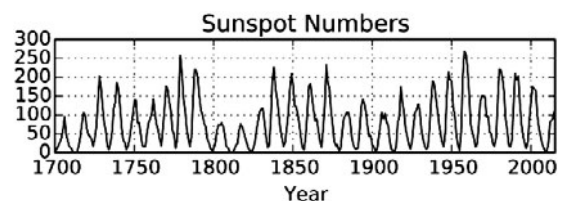


図 1 1700 年から 2013 年までの太陽黒点数。

$$x(t) = 1.19x(t-1) - 0.456x(t-2) + 0.243x(t-9) + \varepsilon_t \quad (6)$$

となる。このモデルから、太陽黒点数の時系列データは、時間遅れ項 $\{x(t-1), x(t-2), x(t-9)\}$ によって現在の状態が記述されていることになる。時間依存性を持つデータのモデルには多くの場合 $x(t-1)$ の項が含まれる。また、時間変化が有限範囲に収まるような定常的データでは、以前の時間変化を打ち消すために逆符号の時間変化が必要であり、それが $x(t-2)$ の項の働きであると考え、ここでは $x(t-9)$ の項の出現が重要である。太陽黒点は約9.5年から12年の周期で増減をすることが知られているが、この9年の時間遅れ項はその情報を与えているものと考えられる。

ここで、比較のために同じデータで AR モデルを構築すると、

$$\begin{aligned} x(t) = & 1.21x(t-1) - 0.424x(t-2) - 0.127x(t-3) \\ & + 0.134x(t-4) - 0.0692x(t-5) + 0.0222x(t-6) \\ & + 0.365x(t-7) - 0.0516x(t-8) + 0.255x(t-9) + \varepsilon_t \end{aligned} \quad (7)$$

となる。このモデルでは1から9までのすべての時間遅れ項が含まれており、モデルパラメータの情報のみから、どの時間遅れ項が周期を表すかを判断するのは容易ではない。

時系列データの性質を理解する上でその周期の推定は極めて重要である。周期の推定法としては、スペクトル推定、自己相関関数、ウェーブレット変換などフーリエ変換やその一般化に基づく手法が広く用いられている。しかし、通常われわれが用いる時系列データは観測ノイズを含む。そのため、多くの場合、それらの手法では正確な周期の特定は困難である。また、複雑な振る舞いを見せる時系列データには、通常複数の周期が存在する。上述した手法で、仮に複数の周期が特定されたとしても、周期どうしが持つ関係を知ることはできない。RAR モデルは、広く用いられている AR モデルを情報理論的に洗練したモデルであり、選ばれる時間遅れ項は時系列データに含まれる周期に対応する。そのため、正確な周期を特定することが可能となる[8]。

ここで、統計モデルにはガウシアン雑音の項が含まれるが、その統計モデルは偶然得られたものではなく、時系列データの持つ性質をきちんと反映したものであることを強調しておきたい。例えば、

$$x(t) = 1.01x(t-1) - 0.61x(t-3) + 0.11x(t-6) + \varepsilon_t \quad (8)$$

から生成した時系列データがあるとする。このモデルが生成する時系列データ10000点から RAR モデルを構築すると、

$$x(t) = 1.01x(t-1) - 0.612x(t-3) + 0.107x(t-6) + \varepsilon_t \quad (9)$$

となり、ほぼもとのモデルが再現される。このようにモデ

ル中には不確定な雑音が含まれるにもかかわらず、時系列データのみを用いて、その時系列データを生成した統計モデルをほぼ正確に同定することが可能である。

また、RARモデルは多変数時系列データのモデリングにも容易に拡張することができ、二体の相互相関を越えたシステム全体の変数間の相関をうまく取り扱うことができることもわかっている。

3.3.4 モデルのネットワーク化

本小節では、時系列データから構築された RAR モデルをネットワークとして可視化する方法として、中村と谷澤によるアイデアを紹介する[1]。

RAR モデル

$$x(t) = a_0 + a_{l_1}x(t-l_1) + a_{l_2}x(t-l_2) + \dots + a_{l_m}x(t-l_m) + \varepsilon_t \quad (10)$$

は、現在の値を表す項 $x(t)$ がいくつかの時間遅れ項 $x(t-l_1), x(t-l_2), \dots, x(t-l_m)$ によって決まっていることを表している。この関係性を表現するため、各項をネットワークノードと考え、各時間遅れ項から現在の項に向かう矢印を引く。(図2参照。)

また、モデルの各項のパラメータについては、絶対値が大きい(小さい)場合、その時間遅れ項が現在に及ぼす影響が大きい(小さい)と考えられる。したがって、ネットワーク化の際には、絶対値の大きな(小さな)パラメータに対応する時間遅れ項(ノード)は現在の項(ノード)の近く(遠く)に置くことが自然であろう。パラメータ値を距離に変換するために、パラメータ値の集合 $\{a_{l_1}, a_{l_2}, \dots, a_{l_m}\}$ をベクトルと見なし、線形代数でよくやるように、現在時刻を表す「方向」と l_j の時間遅れの「座標軸」との間の「角度」 θ_j を

$$\theta_j = \arccos\left(\frac{a_{l_j}}{\sqrt{a_{l_1}^2 + a_{l_2}^2 + \dots + a_{l_m}^2}}\right) \quad (11)$$

と考え、ノード $x(t)$ とノード $x(t-l_j)$ 間の距離を $d = |\tan\theta_j|$ により定義する。この距離は、 $a_{l_j} \rightarrow \infty$ のとき $d \rightarrow 0$ 、 $a_{l_j} \rightarrow 0$ のとき $d \rightarrow \infty$ となる望ましい性質を持っている。例として、前節の最後で取り上げたモデル

$$x(t) = 1.01x(t-1) - 0.61x(t-3) + 0.11x(t-6) + \varepsilon_t \quad (12)$$

の場合、距離 $(d_1, d_3, d_6) = (0.614, 1.67, 10.7)$ となる。

最後に、距離の入ったネットワークにおいては、ノード間の直接距離と間接距離を考慮しなければならない。例えば、式(12)で表されるモデルの場合、時間遅れ6の項は現

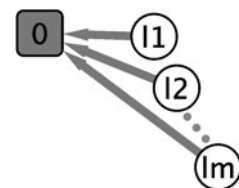


図2 現在の項 $x(t)$ (四角のノード0) に向かって、RAR モデルに表れる各時間遅れ項 $x(t-l_j)$ から矢印を引く。

時刻項と距離 10.7 で直接つながっているが、時間遅れ 3 の項を介して、距離 $1.67 + 1.67 = 3.34$ で間接的にもつながっており、この間接距離 3.34 は直接距離 10.7 よりも小さい。したがって、時間遅れ 6 の項は時間遅れ 3 の項を介する間接接続により現時刻項とつながっていると考えのほうがよい。これらをすべて考慮し、最大時間遅れ 6 までのすべてのノードのつながりを表すと、図 3 のようになる。このネットワーク構造から基本周期 6 の中に時間遅れ 3 の内部構造が見られることがわかる。

同様に、式(6)でモデル化された太陽黒点数の時系列データをネットワーク化したものが図4である。基本周期9の中ですべてのノードが直列につながっており素直な周期運動であることがわかる。

3.3.5 ダイナミクスとネットワークトポロジー

本小節では、時系列データの動的性質がネットワークトポロジーにどのように反映されるのかについて、具体的な例をもとに述べる。

図5は、1982年3月21日に発生した北海道浦河沖地震の地震動を表すデータである。データ点は、0.02秒毎に測定され、全部で2500個の時系列データとなっている[9]。ここでは、図中に示された3つのRegionに注目する。Region 1は初期微動、Region 2は本震、Region 3は後震である。地震動はいくつかの地殻振動が複雑に合成されたものであることが知られ、その周期を正確に取り出すことは容易では

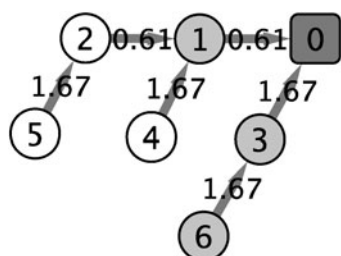


図3 式(12)で表されるモデルをネットワーク化したもの。灰色のノードはモデルの中に現われる項を表す。

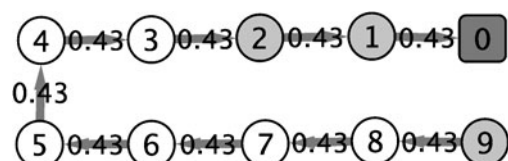


図4 式(6)でモデル化された太陽黒点数の時系列データをネットワーク化したもの.

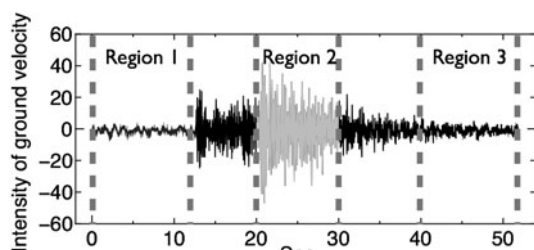


図5 1982年浦河沖地震の地震動。横軸は発生後の秒数、縦軸は地表面の速度である。

ない。

図6はこの地震動の各RegionからRARモデルを構築し、それをネットワーク化したものである。Region 1の初期微動を現わすネットワークは単純な直列形をしておらず、現在時刻に直接接続される項と時間遅れ4（これは0.08秒の時間間隔に相当する）の項に接続された上で現時刻項に接続される構造が連結されていることがわかる。これは、複数の地殻振動が複合して発生する地震発生初期の地震動の複雑さを表すものであると考えられる。そこから時間が経過し地殻構造全体の振動を伴う大きな本震が始まると、その振動は直列構造を取りはじめ、最後の段階では単純な二重直列構造へと落ち着いていく。

ここで参考のために、Region 1 から Region 3 までの RAR モデルはそれぞれ

Region 1 :

$$\begin{aligned} x(t) = & -0.2203 + 0.3652x(t-1) + 0.2811x(t-3) \\ & + 0.4054x(t-4) + 0.1378x(t-6) - 0.1496x(t-7) \\ & - 0.2024x(t-8) + 0.1442x(t-11) - 0.1803x(t-12) + \varepsilon_t, \end{aligned} \quad (13)$$

Region 2 :

$$\begin{aligned} x(t) = & -1.1629 + 1.4154x(t-1) - 1.7348x(t-2) \\ & + 1.2533x(t-3) - 0.8751x(t-4) + 0.1473x(t-5) \\ & - 0.1657x(t-7) + \varepsilon_t, \end{aligned} \quad (14)$$

Region 3 :

$$\begin{aligned} x(t) = & -0.4880 + 0.8847x(t-1) - 0.6488x(t-2) \\ & + 0.2064x(t-3) + 0.2648x(t-4) - 0.1846x(t-5) \\ & + 0.1881x(t-6) - 0.1133x(t-12) + \varepsilon_t \end{aligned} \quad (15)$$

となるが、これらのモデル式のみから、ネットワークポロジに見られる、データが内包する力学的性質の時間的な変化を洞察することはそれほど容易ではないと思われる。

このように、振動状態の「複雑さ」がネットワークトポロジーの「複雑さ」となって直観的に表現されることがこ

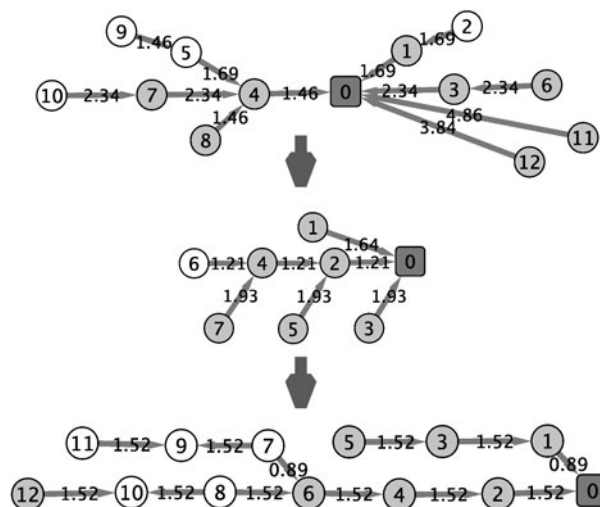


図6 図5の地震動から構築されたRARモデルを本節の方法によってネットワーク可視化したもの。上から順にRegion 1, Region 2, Region 3のネットワーク化である。

の方法の大きな利点である。

3.3.6 プラズマ乱流現象への応用

この方法のプラズマ乱流現象への応用例として、抵抗性ドリフト波乱流数値シミュレーション[10]の解析結果について述べる。軸方向に一様磁場のかかった円筒形プラズマにおいて、密度勾配によって駆動される不安定性をシミュレーションした時の静電位揺動の時間変化を解析対象データとしている(図7参照)[11]。

このシミュレーションデータを上述の方法でRARモデル化すると

$$x(t) = 0.5021x(t-1) - 0.4728x(t-2) + 1.152x(t-5) + 0.2992x(t-8) - 0.4697x(t-10) + \varepsilon_t \quad (16)$$

となる。ここでは、1の時間遅れは0.05[ms]に対応している。このデータは特徴的な周期として約0.25[ms]を持つことがわかっているが、確かにRARモデルでは非自明な5の時間遅れ項として含まれている。

図8は、このRARモデルをネットワーク化したものである。これを見ると現在時刻に強く結びついているのは5の時間遅れ項であることがはっきりとわかる。

3.3.7 最後に

本節では、与えられた時系列データからRARモデルを構築し、その構造をネットワークによって可視化する新しい方法を紹介した。この方法によれば、時系列データの周期の正確な特定のみならず、時系列データが持つ潜在的な力学的性質をネットワークのトポロジーとして直観的に理解することも可能になる。この方法は原理的にはどのような時系列データに対しても用いることができ、非常に応用範囲が広い。その一例としてプラズマ乱流解析への応用例を紹介した。本節の解説を通して、複雑なシステムの動的現象の解明に統計モデルやそのネットワークによる可視化がどのように役立つかについての理解が深まれば幸いである。

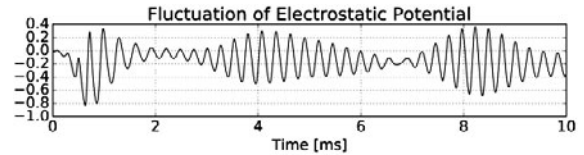


図7 軸方向に一様磁場のかかった円筒形プラズマにおける抵抗性ドリフト波乱流数値シミュレーションから得られた静電位揺動。半径 $r = 0.7a$ (a はプラズマ半径) の位置での時間変化を示す。

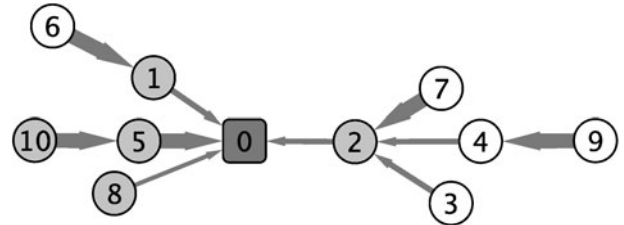


図8 式(16)で表されるRARモデルをネットワーク化したもの。各ノードの数字は時間遅れ、灰色のノードはモデルに現われる項、白色のノードは最大時間遅れ10内だがモデルには含まれていない項を表す。矢印が太いほどノード間の距離が小さい。

参考文献

- [1] T. Nakamura and T. Tanizawa, *Physica A: Statistical Mechanics and Its Applications* **391**, 4704 (2012).
- [2] J.D. Hamilton, *Time Series Analysis* (Princeton University Press, NJ, USA, 1994).
- [3] H. Akaike, *IEEE Trans. Automat. Contr.* **19**, 716 (1974).
- [4] G. Schwarz, *Ann. Stat.* **6**, 461 (1978).
- [5] T. Nakamura *et al.*, *International Journal of Bifurcation and Chaos* **16**, 2153 (2006) and References therein.
- [6] K. Judd and A. Mees, *Physica D* **82**, 426 (1995).
- [7] K. Judd and A. Mees, *Physica D* **120**, 273 (1998).
- [8] M. Small and K. Judd, *Physical Review E* **59**, 1379 (1999).
- [9] T. Takanami and G. Kitagawa, *Methods and Applications of Signal Processing in Seismic Network Operations* (Springer, NY, USA, 2003).
- [10] N. Kasuya, M. Yagi, K. Itoh, and S.-I. Itoh, *Phys. Plasmas* **15**, 052302 (2008).
- [11] シミュレーションデータは九州大学応用力学研究所の糟谷直宏氏より提供を受けた。