Epidemiological Modeling of Knowledge Propagation Represented by Scientific Publications

Daniel Moritz Marutschke

(Degree)
博士（学術）

(Date of Degree)
2014-03-25

(Date of Publication)
2015-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6155号

(URL)
https://hdl.handle.net/20.500.14094/D1006155

Kobe University 神戸大学大学院国際文化学研究科

博士論文

Epidemiological Modeling of Knowledge Propagation Represented by Scientific Publications

(科学論文に基づく知識伝搬の疫学的モデル化)

January 21, 2014

専攻:	グローバル文化
コース:	情報コミュニケーション
氏名:	Daniel Moritz Marutschke
学籍番号:	098C302C
指導教員:	村尾　元　教授

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my immense gratitude to my supervisor Professor Hajime Murao from Kobe University for his encouragement, support, and patience guiding me through this PhD study.

Valuable insights, suggestions, and reviews given by Professors Junya Morishita from Kobe University and Victor V. Kryssanov from Ritsumeikan University has been a great help during my research.

I also would like to thank Professors Min Kan, Hidenari Kiyomitsu, Takeshi Nishida, and Kazuhiro Otsuki for their helpful feedback throughout the doctoral course.

I thank my parents for their support and encouragement, and Tracey Kimmeskamp, who did the final proofreading.

Lastly, I'm grateful to the Yoneyama Rotary Club and JAICA for their financial support which ensured the time needed to complete this study.

ABSTRACT

This doctoral thesis investigates the propagation of complex knowledge—such as scientific research methodologies, algorithms, etc.—represented by the number of scientific paper publications each year. As investigative tools, deterministic differential equation models from epidemiological fields are used. Epidemiology is a well-founded field of research for the spread of diseases and has been successfully applied to tracking epidemics, endemics, and pandemics.

Epidemiological models using differential equations have matured over the last half century in fields such as medicine and biology. Viruses, parasites, and various kinds of contagion agent in a variety of communities, cross-community, vaccination adapted, delayed vaccination campaign, and many more attributes have been researched in-depth. Although the adoption of tracking information diffusion with models from epidemiology dates back to the 1960s, the propagation of more complex knowledge is still insufficiently explored. One of the key challenges is the human factor, as information is transferred from one

individual to another with a highly complex set of properties. Information propagation—such as rumors, expansion of economic fields, opinion, email messages, twitter news, and the like—has been successfully modeled using compartmental models from medical and biological epidemiology. The compartmental models that have proven to be able to track information diffusion are the SIR (Susceptible, Infective, Recovered) and the $SEIR$ (Susceptible, Exposed, Infective, Recovered) models and slight variations thereof. In this paper, the author starts with modeling the number of topic-related keyword propagations in scientific publications using the generic $SEIR$ model. It is apparent that the transfer of knowledge has unique characteristics that cannot be explained by generic epidemiology. Several causally extended models are proposed and tested for their performance to better represent the propagation of knowledge.

The propagation of knowledge represented by scientific publications has been carried out using four data sets. The first data set with 138,303 papers was gathered on the IEEE Xplore Digital Library. Information was accumulated about the paper title, abstract, authors, affiliations, year of publication, and citation count. This allowed for a detailed exploratory analysis of propagation attributes and additional room for refining any algorithms to track the growth of paper numbers.

Three further data sets were acquired using the help of online search engines inside major scientific publication databases. This allowed for extended data-points and some other gathering of semantic information such as cross-referencing the author's affiliation. Another advantage was the ability to effectively conduct full text searches which was previously prohibited due to data size as well as access restrictions. One of the data sets was gathered from CiNii, a search engine for academic information and articles in Japanese with

more than 15 million articles. Another data source was CNKI.NET, China's largest online database of scholarly articles with a total of over 47 million articles. The third source was from the meta search engine Scirus, a database of more than 60 million English publications worldwide.

From a set of prominent science and engineering textbooks, 39 initial keywords—extended to 88 with the Scirus database—from information science and engineering were used to examine the suitability of this approach. Many of the authors' published papers, however, use a set of five keywords from Soft Computing to ensure comparability of the studies. The author will show that categorization of keywords into three source databases is possible using $SEIR$ model parameters and will discuss the limits of epidemiological models and implications for the future of this field.

As one of the main contributions, a modified $SEIRE$ model is proposed to more robustly track knowledge propagation in scientific publications. Two more models, an $SEIRK$ and an $SEIREK$ model have also been proposed and were compared according to performance, causality, and complexity. A deepened understanding of topic propagation in scientific publications is considered the second main contribution.

As a classification and knowledge discovery process, all keywords were categorized using parameter-wise k-means clustering. The classification has proven to be difficult due to the limited number of data points, resulting in a limited number of keywords with sufficient accuracy. This knowledge discovery process could help in determining relationships between topics that have an evolutionary connection.

As suggested by previous research, diffusion of information is also influenced by cultural features. Using a classification process with k-means to be able to assign a propagation

feature to a specific culture, the author analyzed five keywords from Soft Computing in the three data sets—CiNii, CNKI.NET, and Scirus.

As indicated by the above three data sets, culture has an influence on knowledge distribution. Qualitative research of ICT adoption in cultural settings is an active field. Quantitative measures of knowledge propagation connected to cultural settings, however, are still infrequent. From the Scirus database, a new data set was built with 32 country-affiliation, each with 22 fields of knowledge. Using Principal Component Analysis, further trends of culture and knowledge was investigated.

Finally the Scirus database was used to compare the $SEIR$ and $SEIRE$ model with regards to their coefficient of determination (adjusted $\bar{R}^2$), their Basic Reproductive Rate (R_0), as well as the residuals distribution of the model fits. Both models present good performance in tracking knowledge propagation in scientific publications. The $SEIRE$ model showed a more consistent fit, especially with slow propagation and a more randomized distribution of the residuals. From a causality standpoint, this can be attributed to researchers that have acquired knowledge and influence the community without necessarily publishing several papers containing the same knowledge.

CHAPTER 1

INTRODUCTION

Information propagation occurs in different forms. It can take place in communication and interaction between humans and other humans, human-computer interaction, or software only. All of these elements, however, have a common denominator, that at some point there is a human link to receive, process, and forward information. This introduces processes that are difficult to predict or model entirely. Even less complex processes have to be modeled significantly simpler than their real life counterparts.

Knowledge or information has to be defined in context, as it is not clear depending on what aspect of communication the focus is pointed. Rumors can be seen as information, or as a piece of information of a larger context (Dickinson and Pearce, 2003; Dietz, 1967; Okada et al., 2013; Pittel, 1987; Shirai et al., 2012). A piece of news can be seen as information, as well as blog entries, or in general some verbal or written information.

Knowledge on the other hand consists of more complex components. It can be a procedure that can be trained, some methodology used to solve a problem, or simply some piece of information as above. As the definition can differ to a rather large extend, the meaning of knowledge will be explained in detail in this research. Thompson et al. (2006) try to clarify concepts in knowledge transfer. Their conclusion is an apparent inconsistency in terms and definitions regarding knowledge transfer, innovation diffusion, knowledge diffusion, etc. According to their literature review, quite some confusion and mislabeling has occured in this area. This research aims to avoid this confusion.

Knowledge as considered in this thesis represents a research method, methodology, process, algorithm, or the like that scientists use in their research and development. For analytical reasons—including the ease of keyword selection and their distinctness—the field of interest for this research is from science and engineering. A few short examples (methods, procedures, and algorithms) that are well known are ANOVA, Principal Component Analysis (PCA), Fuzzy Logic, Very Large Scale Integration (VLSI), Natural Language Processing, Pattern Recognition, etc.

The principle of information transmission from individual to individual has similar foundations and characteristics. A piece of information, a rumor, some procedure to learn, or otherwise is *taught* to another person. This person propagates this information depending on the community in which it is transferred, the trustworthiness of the individual, the usefulness of the piece of information, complexity of the piece of information, and other related factors. Characteristics change, some are prominent, others are not, contingent on the type of information that is being spread. Taking this into consideration, propagation features can be determined in a community. The community itself may show elements

that influence the propagation. Homogeneity, heterogeneity, size, location, and related attributes can be a factor in assessing information diffusion (or other contagion agent, for that matter).

In this research, propagation of knowledge is tracked via the number of published papers in scientific communities related to a specific topic, reflecting the diffusion process. The number of published papers per year containing a given keyword is gathered and compounded as a new data set. This new data indicates statistics related to the actual knowledge-holders (people who have acquainted themselves with a topic). The propagation is both influenced by the number of papers published and the people teaching and spreading the topic without publishing. This particular characteristic will be addressed with a modified epidemiological model in the following sections. The classic model used to track diseases does not consider this attribute and has some weaknesses in tracking the publication increase correctly as well as causality implications.

1.1 Approaches for Modeling Knowledge Diffusion

As the nature of information diffusion is not fully understood, there are several different approaches to simulate, track, or model propagation.

One approach that is also considered in this research is an epidemiological one, that takes the basic idea of transfer of contagion agent from biology, i.e. viruses and the like, to bring to the transfer of ideas or more complex knowledge. These themselves are di-

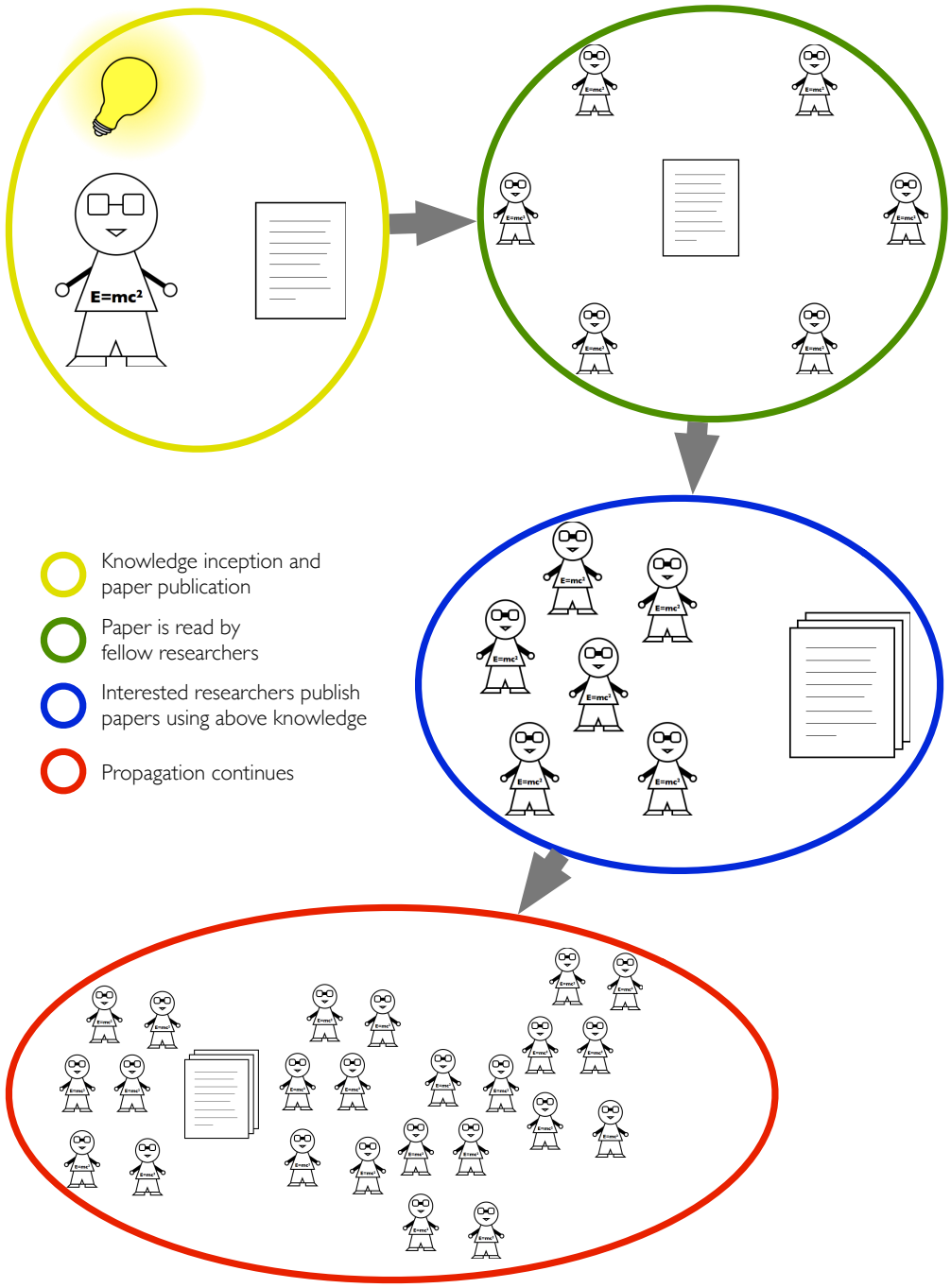


Figure (1.1) Conceptual representation of knowledge propagation by scientific publications.

vided into two sub parts, one statistical modeling and one with compartmental differential equations. The latter, deterministic approach will be described in detail, several models are compared, and a novel construct is introduced to better fit the need for specific knowledge transfer in the case of scientific paper publication (Figure 1.2 shows qualitatively the number of papers over time).

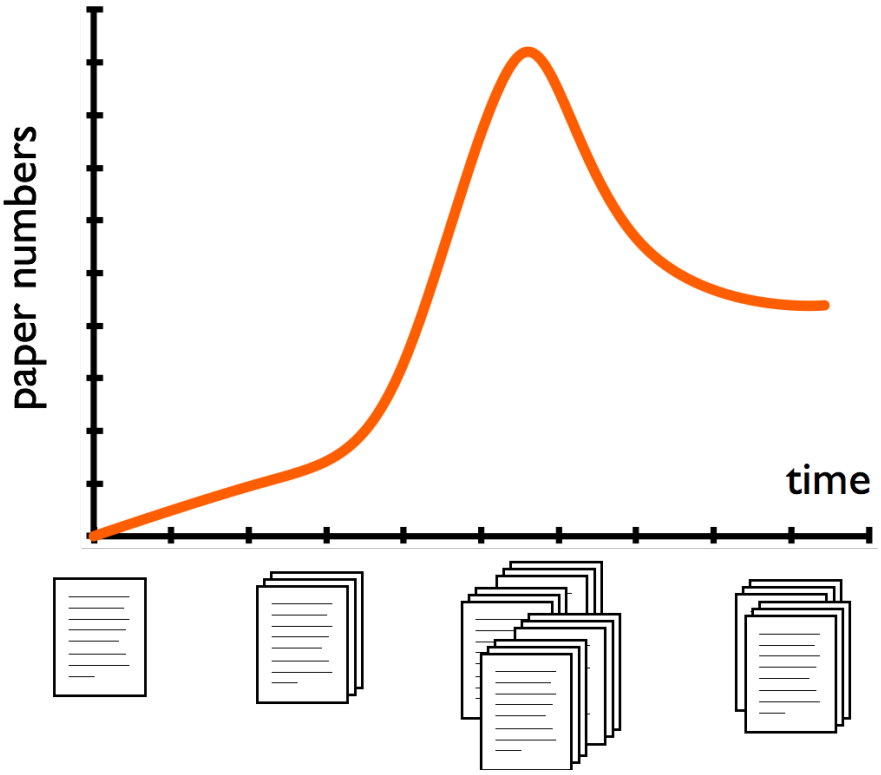


Figure (1.2) Qualitative example of a propagation curve in scientific paper publications.

Related to the first approach, agent simulations can be done of the same compartmental epidemiological models. Agent simulations can add randomness to the propagation process that ordinary differential equations cannot. Also when considering more intricate

interactions between individuals, which is to be expected when modeling human to human interaction with highly complex information exchange and processes, agent simulations can address these obstacles.

Another, still minor but growing concept of modeling knowledge transfer is destination networks (Baggio, 2006; da Fontoura Costa and Baggio, 2009). This method emphasizes the connectivity, partnership, and its force of destination in network structures. This approach is of interest when considering actual individuals traveling and sharing the knowledge on a specific destination. It is a very long-term and relatively slow transformation and not ideal for the research at hand. However, in the early phase of pre-publication stages, this network-based approach might prove to be an element for improvement and will be further investigated. A similar procedure was done by Cowan and Jonard (2001), where agents holding knowledge spread it to other agents on a network. The finding is naturally, that knowledge spreads fastest in compact communities, resembling a "small world" type network.

Cointet and Roth (2007) simulate on two real-world social networks the transmission mechanisms and conclude that the classical stylized network model might not be efficient as knowledge diffusion models.

1.2 Compartmental Epidemiological Models

Compartmental models have the advantage that the transition process is comprehensive and the states are clearly defined. This is of course important for understanding the basic principle behind the transfer process. Figure 1.3 shows three different propagations

graphically with different population sizes (*SIR* and *SEIR* model).

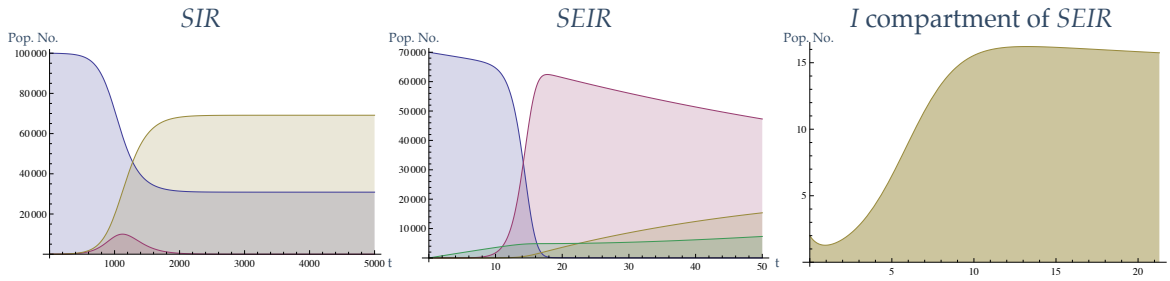
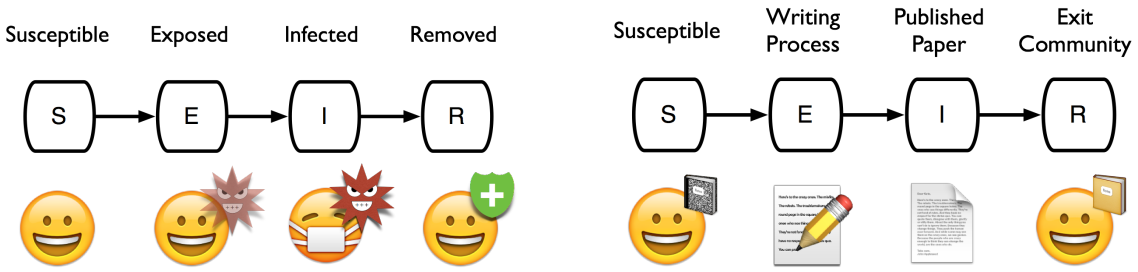


Figure (1.3) Examples of simulations of compartmental deterministic models (*SIR* and *SEIR* models).

The understanding of information diffusion in society has intrigued fields from sociology, including economy and politics (Dietz, 1967; Rhodes et al., 1997; Romero et al., 2011). The subfields in themselves present unique characteristics and properties that in some cases define a new model or approach to tracking correctly. The domains are well understood and they share a common basis, both conceptually and mathematically. And while the mathematics behind these bases are well-established and validated using real data, the understanding of complex human to human information passing is still limited.

One step in deepening the understanding of knowledge propagation is the examination of unique aspects of the diffusion. Nonetheless the fundamental understanding of epidemiological processes is necessary for knowledge diffusion. The basis of this approach has been shown in this research with the usage of the classic *SEIR* model. It is comprehensible that each knowledge as a contagion agent has unique characteristics, similar to different types of diseases. The author believes that an understanding of the fundamental



(a) Compartmental model in medical epidemiology.

(b) Compartmental model in scientific paper publication

Figure (1.4) Illustration of the causal transfer of medical compartmental models to scientific publications and their meaning.

uniqueness of knowledge propagation is necessary to proceed to more complex models. Currently, complex models only adapt to very specific types of knowledge. Without this kind of fundament, the adapted models are prone to lack universal application and the process of trying to fit a certain type of knowledge could mean the analysis of fundamental differences in every case. As one of the main contributions during this research (section 1.4), the author proposes a modified $SEIRE$ model that has a loop back from the removed compartment R to explain knowledge-holders who act in the background (propagate knowledge) without publishing papers, but with the possibility to return to the paper writing process (sections 1.4 and 6.4).

The idea of knowledge transfer using compartmental models from epidemiology was first published by Goffman (Goffman and Newill, 1964). Papers followed, extending the simple epidemiological approach to fit the spread of rumors, news, consumer recognition of a product, and ideas (Daley and Kendall, 1964; Dickinson and Pearce, 2003; Leskovec

et al., 2005; Phelps et al., 2004).

The basic premise set here is the spread of a piece of information, transferred from one person to another in academic fields. Although all forms of information propagation share certain basic behaviors, scientific communities undergo fast changes in their topics. Simple structures of rumors spread are one of the well-researched areas, others have broad recognition but have limits in explaining the whole complex problem. One of these fields is marketing using word-of-mouth, which normally has a goal set to use virus-like promotion of consumer-recognition of a product (Ferguson, 2008).

Concerning this research, the term "knowledge" stands for a research tool such as an algorithm or analysis methodology used in science and engineering. The term "infectious material" refers to the aforementioned knowledge, which an individual (researcher) can transmit to another person. The incubation time can be understood as the time to learn new knowledge or the familiarization with a new research topic.

1.3 Related Research

Involuntary transmission of viruses or parasites as known from medical or biological epidemiology are derived models which follow universal rules of susceptibility, infection, and transfer. Extended characteristics like the behavior of sparks, leading to a Forest-Fire model and the criticality of epidemics are formulated by Rhodes (Rhodes et al., 1997). From these areas, other well-established models were formulated such as SI, SIS, SIR, $SEIR$, $MSEIR$, and other, where S stands for susceptible, I for infectious, R for rejected, E for exposed, and M for immunity from maternity, sometimes until after birth (Heth-

cote, 2000). Dynamics of differential equations in biological, economic, and social sciences are well-established (May, 1976). Mature mathematics in this field are the reason for other fields to emerge, e.g. a $SEIQV$ model—susceptible, exposed, infected, quarantined, vaccinated—to track computer worms with quarantine and vaccination parameters (Wang et al., 2010) or spread in computer networks (Roberto et al., 2005). Noteworthy is the modeling of the forming of extremist opinion by Stauffer with a $GSEF$ model—general, susceptible, excited, fanatic—and Sznajd with a closed community dogma of "united we stand divided we fall" (Stauffer and Sahimi, 2006; Sznajd-Weron and Sznajd, 2001).

1.4 Main Contributions

The two main contributions of this doctoral research are a causally extended deterministic compartmental $SEIRE$ model and a deeper understanding of knowledge propagation in scientific publications.

The $SEIRE$ model as shown in Figure 1.5 was developed to address mechanisms of hidden knowledge-holders, individuals who acquired knowledge with no additional publications but who influence the propagation ($\delta \frac{SR}{N}$), as well as a fraction of researchers who decide to publish a paper again with the same knowledge involved (ϑR). The investigated actual published number of papers is analyzed with the I compartment, marked in orange. Performance and causality comparison as well as formulae and their derivation are written in section 6.4 and an in-depth analysis with the Scirus data set is described in section 6.6.

Essentially all the data sets analyzed during this research have resulted in epidemiological

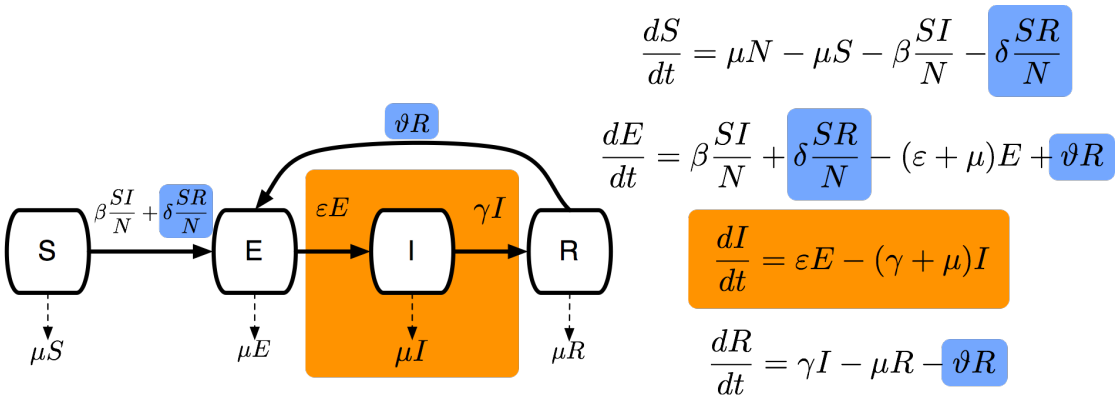


Figure (1.5) *SEIRE* model with corresponding differential equations and relevant extended parameters.

behavior with a timespan of several decades, even for recent advances in science and technology. One consistent observation from the residuals in section 6.6 is an unexplained variance which remains. Fitting the general *SEIR* model to data acquired from scientific publication databases shows high performance suggesting a common causal framework in the propagation process.

The performance of the newly developed *SEIRE* model indicates room for improving classical epidemiological models to fit to specific problems.

1.5 Thesis Overview

Chapter 2 gives a detailed description of epidemiology as understood in medical and biological terms. Based on a literature review, the mathematical foundation and depth of research to date is presented.

Chapter 3 describes previous research and the theoretical foundation of this thesis' main focus. A literature review on past publications, existing models, their contributions, as well as shortcomings are introduced.

Chapter 4 deals with the influence of culture on the diffusion process of information and idea-related topics. Standard models of qualitative assessments of a culture are described. The contribution of quantitative methodologies as in this thesis are discussed to be beneficial to existing cultural models.

As the propagation characteristics have been examined using several data sets, Chapter 5 provides in-depth explanation of the data acquisition procedures, differences in data sets, and the data mining methods.

In the chapter about results and discussions (Chapter 6), the outcomes of the study are described. Potential difficulties are discussed and some ideas for resolution are presented.

The last Chapter 7 sets out the conclusions of the thesis.

CHAPTER 2

EPIDEMIOLOGY

This chapter gives an introduction to epidemiology in its biological origin and provides an overview of its use in today's research field.

The study of epidemiology dates back more than 100 years. Mandal et al. (2011) give an excellent review of the specialization, mathematical modeling possibilities and history of compartmental epidemiological models by means of discussing mathematical models of malaria. Observations and distinctions between an epidemic diseases (one that visits upon a population, spikes inside a population) and endemic diseases (one that resides within a population in an equilibrium) were made as early as Hippocrates, around 400 BC. Mathematical modeling was introduced in the 20th century, becoming more powerful and versatile with the emergence of high performance computers.

"Although some excellent epidemiologic investigations were conducted before the

> 20th century, a systematized body of principles by which to design and evaluate epidemiology studies began to form only in the second half of the 20th century." Chapter 1 – Modern Epidemiology
>
> Rothman et al. (2008)

> "Compared to the classic statistic analysis in epidemic research, employing the well-developed modern theory of dynamic systems and utilizing high-speed computing facilities, epidemic dynamics studies provide deeper understanding of transmission mechanisms and global analysis of transmission dynamics." Basic Knowledge and Modeling on Epidemic Dynamics – Dynamical Modeling and Anaylsis of Epidemics
>
> Ma (2009)

Epidemiology has made major contributions to the understanding of the role of individual lifestyle factors and health. One of the best-known clinical health studies is the connection between cigarette smoke and lung cancer. Epidemiology is dealing with three aspects, an appropriate study design, careful consideration to avoid bias, and the use of suitable statistical methods for analysis (Kleinbaum et al., 2007). Although many aspects of modern epidemiology go back to ancient times, in the 21st century the focus is moving from a *risk factor approach* and from individual lifestyle to include the importance of context (Pearce, 2005; Van den Broeck and Brestoff, 2013). Contexts include population, culture, history, environment, genetic predispositions, etc. Another shift is the problem-based approach to epidemiology, where an issue in a community is identified and addressed. Stehlé et al. (2011) conducted a simulation using high-resolution data from face-to-face communication at a conference, based on Radio-frequency Identification (RFID), using the

SEIR model. The latter study implies the need for a high level of detail that is required to correctly feed computational models with necessary data.

This research is not an *epidemiological study*, e.g. to try to provide accurate answers to questions such as "what is the prevalence of smoking in this district's population?" or "what is the additional risk of liver cancer due to previous Hepatitis B infection?" (Stewart, 2002). This research is about the epidemiological behavior of contagion agent—may they be of viral, bacterial, parasitical response, or rumors, word-of-mouth, information—and using this behavior to understand propagation characteristics of complex knowledge.

2.1 Definition and Background

Epidemiology is the study of disease's dynamics. It tries to find answers to what kind of people get the disease, the influence of population heterogeneity, and a way to counteract the propagation (Stewart, 2002). From these analyses, usually a strategy for managing or preventing established diseases is initiated. This generally involves taking samples of a population.

> "Epidemiology is the study of how often diseases occur in different groups of people and why. Epidemiological information is used to plan and evaluate strategies to prevent illness and as a guide to the management of patients in whom disease has already developed."
>
> Coggon et al. (1997)

Mathematical epidemiological models range from deterministic models and stochastic models, to network models and, recently, multi-agent modeling. Deterministic compart-

mental models are among the most researched and seem to have started at the beginning of the 20th century (Hethcote, 2000). Hamer in 1906 and Ross in 1911 may have been the first to formulate models in an attempt to understand the epidemiological foundation (Hamer, 1906; Ross, 1911a). Papers followed by Ross as well as Lotka (Lotka, 1922; Ross, 1911b). Kermack and McKendrick contributed to the field of epidemiology from 1926 with research on epidemic thresholds (Kermack and McKendrick, 1991; M'Kendrick, 1925). This research area continued to grow, exponentially so from the mid-20th century. The emergence of faster computers helped by allowing more complex modeling (Murillo et al., 2013).

2.2 Types of Models

Within the spread of disease, there are five compartments—M, S, E, I, and R. Figure 2.1 shows the transition from one compartment to another. Newborn babies whose mothers pass antibodies through the placenta to the infants land in the M compartment with passive immunity. They later move to the susceptible compartment S where they can contract the disease. Compartment E denotes the incubation time, the I compartment the infectious stage. After immunization or death, individuals move to the recovered or removed compartment R.

Depending on the type of disease, the incubation time can be minimal, resulting in an SIR model. Or the once-immune individual can contract the disease again after some period, which is described by $SEIRS$, $MSEIRS$, or $SIRS$ models. In cases where the

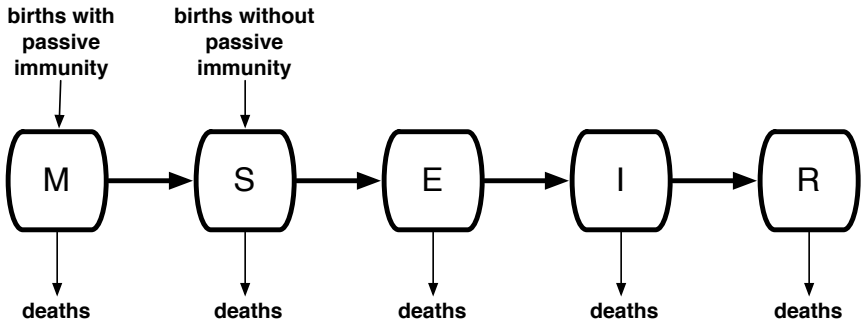


Figure (2.1) The $MSEIR$ model with transitions.

transition ratio α from the R back to the S compartment goes $\alpha \to \infty$, models such as $SEIS$ or SIS can explain the spread.

These models have been researched in constant populations, exponential population dynamics, homogeneous and heterogeneous populations, and with the impacts of vaccination on the propagation process in various population sizes, epidemics, endemics, and pandemics (Hethcote, 1976, 1989; Hufnagel et al., 2004; Li et al., 1999; Mena-Lorcat and Hethcote, 1992; Prematunge et al., 2012).

2.3 Deterministic Models

Deterministic epidemiological compartmental models are represented by ordinary differential equations. The number of differential equations in the system depends on the number of compartments. It can range from a two-equation model to a six-equation model or more, if e.g. age groups are influencing the system.

Classical models such as the SIR (Equations 2.1a–2.1c) model or the $MSEIR$ (Equa-

tions 2.3a–2.3f) model have the following structure.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \beta\frac{SI}{N} \tag{2.1a}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta\frac{SI}{N} - \gamma I \tag{2.1b}$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \tag{2.1c}$$

The Basic Reproductive Rate for the *SIR* model is given by

$$R_0 = \frac{\beta}{(\gamma - \mu)}, \tag{2.2}$$

where β is the contact rate and $1/(\gamma - \mu)$ the average death-adjusted infectious period (Hethcote, 2000). The Basic Reproductive Rate, in scientific papers often called *Basic Reproduction Number*, is a measure of the number of infections produced, on average, by an infected individual in the early stages of an epidemic, when virtually all contacts are susceptible

$$\frac{\mathrm{d}M}{\mathrm{d}t} = b(N - S) - (\delta + d)M \tag{2.3a}$$
$$\frac{\mathrm{d}S}{\mathrm{d}t} = bS + \delta M - \beta\frac{SI}{N} - dS \tag{2.3b}$$
$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta\frac{SI}{N} + (\varepsilon + d)E \tag{2.3c}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - (\gamma + d)I \tag{2.3d}$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - dR \tag{2.3e}$$
$$\frac{\mathrm{d}N}{\mathrm{d}t} = (b - d)N \tag{2.3f}$$

The Basic Reproductive Rate for the $\boldsymbol{MSEIR}$ model is given by

$$R_0 = \frac{\beta\varepsilon}{(\gamma + b)(\varepsilon + b)}, \tag{2.4}$$

where β is, as above, the contact rate, $1/(\gamma+b)$ the the average infectious period adjusted for population growth, and $\varepsilon/(\varepsilon + b)$ the fraction of exposed people surviving the latent class $\boldsymbol{E}$ (Hethcote, 2000).

Figure (2.2) Illustration of the stages in the *SIR*, $\boldsymbol{SEIRS}$, and $\boldsymbol{MSEIR}$ models.

The transition parameters are described for the $\boldsymbol{MSEIR}$ model.

The birth rate into the susceptible class is bS and corresponds to the newborns whose mothers are susceptible. The number of newborns into the passive immune compartment $\boldsymbol{M}$ are given as $b(N - S)$.

Parameter d denotes the death rates and corresponds with each class as $\mathrm{d}M$, $\mathrm{d}S$, $\mathrm{d}E$, $\mathrm{d}I$, and $\mathrm{d}R$, respectively. The transfer rate from the passively immune class to the susceptible class is δM. The transfer from the exposed compartment $\boldsymbol{E}$ into the infective class $\boldsymbol{I}$ is given by εE and the recovery rate from the infectious compartment is denoted

by γI.

Many adapted models have been proposed to address specific diseases such as the varicella-zoster virus, a herpesvirus (Schuette, 2003). The author modified the $\boldsymbol{MSEIR}$ model by adding two compartments $\boldsymbol{W}$ and $\boldsymbol{Z}$. The $\boldsymbol{W}$ compartment stands for individuals with weak immunity and the possibility to reenter the $\boldsymbol{R}$ compartment with sufficient contact to the virus[1]. The $\boldsymbol{Z}$ compartment are individuals where this boosting does not happen and these individuals contract the virus.

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mu(N - N - S) - \delta M - \mu M \tag{2.5a}$$
$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu(M + S) + \delta M - \beta\frac{I + \rho Z}{N}S - \mu S \tag{2.5b}$$
$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta\frac{I + \rho Z}{N}S - \varepsilon E - \mu E \tag{2.5c}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - \gamma I - \mu I \tag{2.5d}$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I + \beta\frac{I + \rho Z}{N}W + \eta Z - \kappa R - \mu R \tag{2.5e}$$
$$\frac{\mathrm{d}W}{\mathrm{d}t} = \kappa R - \beta\frac{I + \rho Z}{N}W - \sigma W - \mu W \tag{2.5f}$$
$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \sigma W - \eta Z - \mu Z \tag{2.5g}$$

Due to a number of additional parameters and average survival ratios in the incubation time, the reproduction number is more complicated and not as intuitive to derive than from SIR, $\boldsymbol{SIS}$, SEIR, and similar models. In the case of Schuette (2003), the reproduction number R_0 is as follows:

[1] This hypothesis is called *boosting* and is not widely acknowledged. Schuette (2003) uses this mechanism to control the reactivation of latent varicella-zoster virus.

$$R_0 = \beta\frac{\varepsilon}{\bar{\varepsilon}}\frac{1}{\bar{\gamma}} + \rho\beta\frac{\varepsilon}{\bar{\varepsilon}}\frac{\gamma}{\bar{\gamma}}\frac{\kappa\rho}{\bar{\kappa}\bar{\rho}\bar{\eta} - \kappa\rho\eta}$$

The contact rate is given as β, the average part that survive the incubation period as $\varepsilon/\bar{\varepsilon}$, the average infection period as $1/\bar{\gamma}$, and ρ as the ratio of the infectiousness from those in compartment Z to compartment I. Each μ-adjusted parameter is written with a bar over the variable (e.g. $\varepsilon + \mu = \bar{\varepsilon}$, $\kappa + \mu = \bar{\kappa}$, etc.). All transitional parameters are given as δM, εE, γI, κR, σW, and νZ.

CHAPTER 3

KNOWLEDGE PROPAGATION

The term *knowledge* has no unified definition which leads to different interpretations depending on the study. In this thesis, *knowledge* is understood to be research methodologies, algorithms, and the like which scientists use in their research. One of the goals of this thesis is to model the propagation of this knowledge. As a representation of the transfer of knowledge and the spread thereof, scientific publications are text-mined for keywords corresponding to a research methodology or algorithm.

As in epidemiology, this research is assessing the study population of scientific publications to make inferences about the target population, is this case the research community (Van den Broeck and Brestoff, 2013).

As an example a researcher discovers a novel method of solving a problem and publishes a paper about that method. Other researchers read the paper and study the new

methodology to implement in their own research and as a consequence also write a paper using the same method. This propagation hypothesis leads to characteristics found in epidemics.

3.1 Background

The idea of knowledge transfer using compartmental models from epidemiology was first published by Goffman (Goffman and Newill, 1964). The paper history continues with works from Allen (1982); Bartholomew (1973); Cavalli-Sforza (1981); Funkhouser and McCombs (1972); Karmeshu and Pathria (1980).

Recent publications are extending the simple epidemiological approach to fit the spread of rumors, news, consumer recognition of a product, and ideas (Dickinson and Pearce, 2003; Dodds and Watts, 2005; Leskovec et al., 2005; López-Pintado, 2008; Phelps et al., 2004; Tabah, 1999; Watts, 2002; Watts et al., 2007).

3.2 Previous Work

This section lists and explains models that have the closest relationship to this research.

The first model is the classic $SEIR$ model that has been used unaltered to track even complex knowledge as in the current analysis. It consists of four compartments, each represented by an ordinary differential equation with transition parameters and relationships. The S compartment represents the susceptible compartment, hence people who can get infected (Equation (3.1a)). The E compartment holds individuals that have contracted a

contagion agent but are not yet symptomatic (Equation (3.1b)). The I compartment is the one with the infectious individuals (Equation (3.1c)). The R compartment resembles the removed individuals, that is the ones that have become immune, deceased, or otherwise removed from the community (Equation (3.1d)).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \mu S - \beta \frac{SI}{N} \tag{3.1a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta \frac{SI}{N} - (\varepsilon + \mu)E \tag{3.1b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - (\gamma + \mu)I \tag{3.1c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R \tag{3.1d}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}E}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} = 0 \tag{3.1e}$$

From Equation (3.1e), the total population N equals to the sum of S, E, I, and R.

Another relevant model is a derivative of the original *SEIR* model with an additional Z compartment to simulate stiflers or actively rejecting individuals. The idea that Bettencourt et al. (Bettencourt et al., 2006) present is that in communities with competing ideas, there exist individuals who actively make negative publicity against specific knowledge in favor of the propagation of another idea. This model has been accurate in tracking the beginning of the propagation of a methodology from physics in three countries, USA, Japan, and the

USSR (Equations (3.2a) to (3.2d)).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta\frac{IS}{N} - b\frac{SZ}{N} - \mu S \tag{3.2a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = (1-p)\beta\frac{SI}{N} + (1-l)b\frac{SZ}{N} - \rho\frac{EI}{N} - \varepsilon E - \mu E \tag{3.2b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = p\beta\frac{SI}{N} - \rho\frac{EI}{N} + \varepsilon E - \mu I \tag{3.2c}$$

$$\frac{ZI}{\mathrm{d}t} = lb\frac{SZ}{N} - \mu Z \tag{3.2d}$$

The Basic Reproductive Rate for the $\boldsymbol{SEIZ}$ model is given by

$$R_0 = \frac{\beta(\varepsilon + p\mu)}{\mu(\varepsilon + \mu)}, \tag{3.3}$$

where β is the contact rate from $\boldsymbol{S}$ to $\boldsymbol{I}$ compartment, p the transition probability $S \to I$ given contact with adopters of the idea, and $1/\varepsilon$) average idea incubation time and $1/\mu$ the average idea-lifetime.

Gurley and Johnson (2011) proposed a minor extension of the classic $\boldsymbol{SEIR}$ model using constants to better fit the propagation of subfields in economics. In their paper, the fields of economics are tracked with high statistical significance, resulting in the ability to detect subfields that will expand rapidly or are more likely to collapse. The model is based on a citations count but information adoption or the propagation of ideas in general is not analyzed.

Kiss et al. (2010) investigate the mapping of networked information diffusion with the $\boldsymbol{SEI}$ (Susceptible—Exposed—Infectious) and $\boldsymbol{SI}$ (Susceptible—Exposed) model.

(Yuan and Chen, 2008) have modeled virus spreading on computers with a multi-state

antivirus, latent period, and point-to-group propagation with an *e-SEIR* model.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \nu S - \rho_{SR} S - \mu S \tag{3.4a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \nu S - m_E \alpha E \tag{3.4b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \alpha E - (\gamma + \mu) I \tag{3.4c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \rho_{SR} S + \rho_{ER} E + \gamma I - \mu R \tag{3.4d}$$

Realtime immunization is denoted as ρ_{SR}, the impact of cleaning the virus and immunizing the nodes in the latent period as ρ_{ER}, and μ is the influence of quarantine or replacement. Parameter ν is the transition rate from S to E, α the transition rate from E to I compartment, γ the recovery rate from I to R, and r the averaged number of neighbor nodes (directly connected with a infectious node). The state transition rate $1/t_E$ is given as $m_E \alpha = \alpha + \rho_{ER} + \mu$.

The reproduction number is given as follows:

$$R_0 = \frac{r\mu}{m_E(\rho_{SR} + \mu)} \tag{3.5}$$

Yuan and Chen (2008) conclude that in long-term virus propagations, viruses were spreading with $R_0 > 1$ and died out with $R_0 < 1$. Pre-immunization could help reduce follow-up costs significantly.

Recently, Twitter[1] has become the center of attention for modeling the spreading of information, mainly using the *SIR* model due to the similarity with diseases with short

[1] Twitter is an online social networking service and the worldwide most used microblogging site (https://twitter.com).

incubation time. Noticeable are the character limitations to 140 characters per message (Tweet) and coupled with this the ease of reading, understanding, and re-tweeting (forwarding) a message.

Because of the nature of fast, almost instantaneous forwarding of relevant messages, there is no E compartment, which removes the notion of incubation time in the process. The result is a modified *SIR* model by Abdullah and Wu (2011) to fast-track the message spread (Equations (3.6)). Other research with transmission on Twitter include Romero et al. (2011), Shirai et al. (2012) and Okada et al. (2013).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI + \mu I \tag{3.6a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I \tag{3.6b}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \tag{3.6c}$$

Note that μ is defined as the birth rate in Abdullah and Wu (2011) opposite to most other works (including this PhD thesis), where μ is given as the death-rate.

Shirai et al. (2012) and Okada et al. (2013) use the same set of differential equations as in Equations (3.7). Their approach is to identify the influence of rumors or false information in contrast to correct information. The dynamic, however, differs from the causality of the classical *SIR* model in that it incorporates two states of incubation, and two states of adoption. The naming of the R_{get} and R compartments are unfortunate as they represent the exposed compartment to correct information (R_{get}) and the adoption thereof (R), respectively. From a rumor-perspective, on the other hand, they can be viewed as removed

from population.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{F}{N}IS - \frac{F}{N}RS \tag{3.7a}$$

$$\begin{aligned}\frac{\mathrm{d}I_{get}}{\mathrm{d}t} &= (1-\rho_{(S\to I)})\frac{F}{N}IS \\ &\quad - \rho_{(I_{get}\to I)}\frac{F}{N}I_{get}I \\ &\quad - \frac{F}{N}I_{get}R\end{aligned} \tag{3.7b}$$

$$\begin{aligned}\frac{\mathrm{d}I}{\mathrm{d}t} &= \rho_{(S\to I)}\frac{F}{N}IS \\ &\quad + \rho_{(I_{get}\to I)}\frac{F}{N}I_{get}I \\ &\quad - \frac{F}{N}IR\end{aligned} \tag{3.7c}$$

$$\begin{aligned}\frac{\mathrm{d}R_{get}}{\mathrm{d}t} &= (1-\rho_{(S\to I)})\frac{F}{N}IS \\ &\quad + (1-\rho_{(I_{get}\to R)})\frac{F}{N}I_{get}R \\ &\quad + (1-\rho_{(I\to R)})\frac{F}{N}I_{get}I \\ &\quad - \rho_{(R_{get}\to R)}\frac{F}{N}R_{get}R\end{aligned} \tag{3.7d}$$

$$\begin{aligned}\frac{\mathrm{d}R}{\mathrm{d}t} &= \rho_{(S\to R)}\frac{F}{N}SR \\ &\quad + \rho_{(I_{get}\to R)}\frac{F}{N}I_{get}R \\ &\quad + \rho_{(I\to R)}\frac{F}{N}I_{get}I \\ &\quad + \rho_{(R_{get}\to R)}\frac{F}{N}R_{get}R\end{aligned} \tag{3.7e}$$

The authors assume that both R states have no influence on the other states[2]. Com-

[2]Although this is contrary to works of Dietz (1967), Pittel (1987), Dickinson and Pearce (2003), and

partments I_{get} and I are exposed to and adoptive of rumors or false information.

The ρ parameters of Equations (3.7) give the infectious probability from compartment to compartment.

In their paper, Shirai et al. (2012) and Okada et al. (2013) analyze seven different rumors and misinformation—regarding Cosmos Oil Company, power saving, iodine (chemical element), Turkey (country), Taiwan, Pokemon, and Fuji Television. The authors used their model to classify rumors into four categories, one-time simultaneous epidemic, one-time separate epidemic, recurring simultaneous epidemic, and recurring separate epidemic.

Xiong et al. (2012) proposed a diffusion model for the purpose of tracking information on microblogging sites using Twitter. The authors, however, use an approach that includes an incubation time, contrary to the research above. The $\boldsymbol{SCIR}$ model (Equations (3.8)) suggested in their paper resembles four states, susceptible, contacted, infected, and refractory. The basis is similar to the $\boldsymbol{SEIR}$ model with the following stages: coming into contact with the infectious material, gaining interest, spreading the information, and then eventually losing interest in it. The model is given as

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -kSI, \tag{3.8a}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = (1-\lambda)kSI - \lambda k(1-\delta)CI - \delta C, \tag{3.8b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \lambda kSI + \lambda k(1-\delta)CI, \tag{3.8c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \delta C, \tag{3.8d}$$

Belen (2008), it can be assumed that rumor dynamics behave differently on fast-paced microblogging sites.

where k is the average contact rate, λ the average infectious rate, and δ the average refractory rate (i.e. individuals lose interest).

No further insight is given regarding the Basic Reproductive Rate in Abdullah and Wu (2011), Shirai et al. (2012), Okada et al. (2013), and Xiong et al. (2012).

One noteworthy research that could deal with evolutionary characteristics of knowledge propagation (especially considering diminishing effects) is a paper by Henrich (2002). He uses the Price Equation[3] to analyze the decrease of craftsmanship and knowledge about herbal medicine in Tasmania after being isolated for over 8,000 years with a relatively small population size[4]. The equation is given as

$$\Delta\bar{z} = \underbrace{\mathrm{cov}(w/\bar{w}, z)}_{\text{Selective Transmission}} + \underbrace{\mathrm{E}((w/\bar{w})\Delta z)}_{\text{Incomplete Inference}}, \tag{3.9}$$

with $\Delta\bar{z}$ being the average change in skill per time-step. The selective transmission $\mathrm{cov}(w/\bar{w}, z)$ represents the influence of culture on the selection process of skills to be passed on to next generations. The last terms refers to errors that occur in the transmission process (resulting in evolutionary changes over generations). Each $w/\bar{w}$ can be understood as the degree to which an individual focuses on learning a particular skill.

[3] The Price Equation is understood to unify models of kin selection (Frank, 1997).

[4] Small in this context refers to a critical number of individuals necessary to ensure enough diversity to help sustain strains of knowledge that is passed on to next generations.

CHAPTER 4

CULTURE AND PROPAGATION

Several papers investigate the adoption of technology and the influence of culture on it (Erumban and de Jong, 2006; Herbig and Palumbo, 1994; Kaba and Osei-Bryson, 2013). Erumban and de Jong (2006) analyze ICT (Information and Communications Technology) adoption across countries using a power-distance dimension, reflecting the power distribution in a country, uncertainty avoidance as a degree of social uncertainty and ambiguity, individualism concerning the relation between individuals in a group, masculinity-femininity as a cultural characterization, long-term orientation related to cultural tradition and its value to the inhabitants. From graphical and regression analysis, Erumban and de Jong conclude that the ratio of ICT adoption is dependent on cultural factors.

Successfully implementing management style, strategy, and firm performance are important in general but often get complicated in multicultural environments (Gales, 2008).

These studies look at the impact of "Cultural Distance" on the ability of organizations to successfully interact. Further research in this direction has been done by Tchaicha (2005). Tchaicha (2005) discusses how to approach cultural differences in an innovative way. Interdisciplinary courses for cultural awareness are being presented to address the intersection of culture, technology, and business practices and how these three influence one another.

The adoption of information and communication technologies (ICT) and the influence of culture upon it are a key focus of much research. Gallivan and Srite (2005) review the literature to IT and culture. Gallivan and Srite (2005) argue that both areas of research have not been cooperatively approached, thus leaving many questions and interactions unanswered. The authors identify areas of improvement and the opportunity for mutual benefit in multicultural environments. See also Perkins and Neumayer (2005) on diffusion characteristics in developed versus developing countries and countries that are "open" to new technologies, which are both supported by this paper.

Especially the power-distance can pose problems in surroundings with a different cultural background. Isaacs (2001) identifies problems in power-distance when the user and expert of an IT system are not in agreement.

Often referred to is the case of computer adoption in countries. Caselli and II (2001) found strong evidence for this case and a higher level of human capital and the openness to manufacturing imports from the OECD. One of the most robust findings according to the study of Caselli and II (2001) was the striving for high educational achievements.

The importance of innovation in economic growth is increasing and the demand for imitation of technology (referred to as "diffusion" in the paper of (Fagerberg and Verspagen, 2002)) is increasing. The authors find that manufacturing has lost much of its growth

dynamic which is reasoned to be related to technological shifts during the last decades.

It is apparent that different cultures have different characteristics in technology adoption, information spread, and knowledge propagation. As previous research shows, in the case of technology adoption there are many indicators for cultural influence on the subject. In the case of knowledge propagation, the attributes of scientific communities as well as cultural aspects come into play.

This research does not attempt to explain the cultural indicators but to see if the characteristics are consistent in research topics in each cultural environment. In the case of Soft Computing this is the case. Soft Computing, especially the fields that are used in this thesis (Fuzzy Logic, Neural Network, Bayesian Network, Genetic Algorithm, and Evolutionary Algorithm), has been successfully applied in numerous applications (Bonissone, 1997). The epidemiological features are shared with all the research topics so far, ranging from computation to statistics to engineering (Marutschke and Murao, 2013).

Although the classification using PCA and k-means clustering was significant for cultural categories, the classification of each keyword (independent from cultural environment) was considered less significant. In this particular case the number of keywords was not sufficient to compensate for the amount of publications, which was a strong identifier. Keyword-wise categorization is one of the main areas for further research and is presumed to give a deeper understanding of the relationship of fields of knowledge (research methodologies) to each other. Relationships could be evolutionary, i.e. one method replacing another or changing the field in which it is applied.

Cultural dimensions as described in Erumban and de Jong (2006) and Smith et al. (2013) could be combined with this research to find parameters of epidemiological or

evolutionary characteristics linked to cultural aspects.

The Hofstede's dimensions of natural culture consist of five dimensions—individualism vs. collectivism as a degree of people's individuality or with regard to group, team, organizational, or community welfare; uncertainty avoidance, the degree to which people avoid or manage uncertainty; power-distance as a balance or imbalance of economical and social stratification; masculinity-femininity, gender roles, equality, and community; Confucian dynamism as short-term, long-term, future-oriented, or present-oriented communities (Taras et al., 2012).

Hall's cultural factors are context and time. Hall presumes that meaning is a function of information and context (as in cultural habits, often outside conscious awareness) and time as culture evolves and adapts over time.

Schwartz' human values is a recent approach to build a cultural framework. It consists of three components—relations between individuals and group, assuring responsible social behavior, and the role of humankind in the natural and social world. These are in the context of seven domains—conservatism, intellectual autonomy, affective autonomy, hierarchy, egalitarianism, mastery, and harmony. These fine-tuned cultural indicators are considered for future research. Parameters extracted from the deterministic models could quantify these dimensions.

Applications of this research could contribute to the detection of cultural hot-zones in which new knowledge spreads particularly fast. Another application would be an alert system for culturally-based scientific weaknesses. Both of these directions could be used to prevent a cultural group from falling behind in the research community. Understanding the connection between the existing cultural models and the epidemiological properties

of knowledge in scientific communities could provide a way of quantifying the qualitative nature of cultural descriptions.

CHAPTER 5

DATA SETS

The course of a propagation of complex knowledge can take decades to reach a peak and several decades can follow where the topic is still actively used. Some of the foundations of computation have even been developed over centuries. Modern computation and its algorithm have mostly been applied since the 1970s. Other basics in research, such as the ANOVA have their time of inception more than eight decades ago[1] but take several decades to develop an epidemic-like expansion. A combination of information diffusion via the Internet with growing computational power, more complex algorithm, and computationally expensive research, these methodologies are finding their way into scientific publicity.

[1] In Fisher (1918) the term "Analysis of Variance" was coined and three years later applied in a paper by Fisher (Fisher, 1921)

The assumption is made that the number of researchers (in this study mostly scientists and engineers) incorporating specific knowledge can be represented by the number of published scientific papers. Papers in section 3.2 make similar assumptions. As the focus of this study is on scientists and engineers in research areas, this approach was chosen, whereas other considerations such as patents could yield more accurate representations for commercial technology propagation and adoption.

To be able to accurately track the propagation of several different keywords, a large data set was necessary. As knowledge in this context is tracked via textual keywords related to a research topic, method, methodology such as an algorithm, whose propagation is mapped via the number of papers published per year mentioning that specific keyword, several tens of thousands of papers were targeted to be collected.

During the course of the PhD study, effectively four different data sets were composed. The initial data set as described in section 5.2.1 was gathered with the abstract of each paper as the main source of information in mind. The keyword search was performed on the offline data set. The three other data sets (sections 5.2.2-5.2.4) essentially equal a full text search using online search engines of major publication databases.

All the data sets used for this research are unadjusted to possible changes in population size. Other studies presume the same (Bettencourt et al., 2006; Gurley and Johnson, 2011; Marutschke and Murao, 2013; Okada et al., 2013; Xiong et al., 2012), but the length of propagation suggests that a growing number of researchers influence the data set. Also adjustments due to technology advances, such as the digitalization of papers, online tools for searching large databases, etc. are not implemented. Some of these effects are discussed in section 6.

Just about any study that involves data is subject to errors of which *Random Error* and *Systematic Error* are shortly discussed (Newman and Newman, 2001). The data might suffer from systematic error, due to reliance on the integrity of the online databases. An attempt has been made to minimize some of these effects by taking 10 random samples[2] of each propagation and running the fitting algorithm for all samples. The random error is judged to be negligible noise as data-points are quite sufficient, especially in the Scirus database (section 5.2.4).

5.1 Data Acquisition

The first data set that was acquired for this research was done in several steps. Considering the enormous dimension of full text papers (usually as PDF files), obtaining that data for offline usage was impractical based on sheer size and problematic considering copyright of the material and the strain on servers of the content providers. However, most online publication databases that list scientific papers for download have preview of information such as title, abstract, year of publication, authors, and other paper related statements that can be accessed without having an account or the credentials to download the paper.

To facilitate the keyword selection and tracking process, scientific papers were judged to be most applicable. A selection of journals, proceedings, and transactions as described

[2]Random samples of the same sample size as each original vector were taken by using Mathematica's RandomChoice which is based on cellular automaton to generate high-quality pseudorandom numbers (http://reference.wolfram.com/mathematica/tutorial/RandomNumberGeneration.html)

in section 5.2 were selected for download. The first step was obtaining a list of the basic relationship of links associated with the paper names. The next step involved extracting all URLs of the aforementioned relationship to associate with a specific paper. The last step was the textual download of each of the link into a separate file. This concludes the raw acquisition process, which was followed by steps of data mining, described in section 5.3.

5.2 Data Sets in Detail

This section describes each of the four data sets in detail. The IEEE Xplore Digital Library differs from the other data sets in that it was acquired for offline use by several crawling algorithms whereas in the other databases the search was done online and the results were extracted via automated algorithms (minor web-crawling).

5.2.1 IEEE Xplore Digital Library

To be able to flexibly track keywords through large text data sets, a data set was built from the IEEE Xplore Digital Library. The data was compiled online using crawling algorithms and filters for information extraction. The total data set consists of 138,303 papers with information about publication affiliation, authors, authors' affiliation, title, abstract, year of publication, and citation count. A collection of 32 journals, transactions, and proceedings was hand-selected to compile the data set, covering a timespan of 73 years from 1939 to 2011. To explore the suitability for future research, the criteria for a paper to be included into this research data set was that the existence of the publisher has a timespan of more than 10 years and that the field of research was related to information science

and engineering.

Two exceptions were made for selecting publications; the Electronics & Communication Engineering Journal with a timespan of six years, due to the high paper count and relevance to this research's field, and the IEEE Transactions on Automation Science and Engineering with nine years of publications. As this research is about the number of papers per year that have been published regarding a specific topic, title and abstract were sought out to test for usability. Due to the limited number of data points, the abstracts with substantially more text-data to process were used.

See Appendices B and A for more details.

5.2.2 CiNii

CiNii is a search engine for academic information and articles in Japan with more than 15 million articles (as of October 2013). This database is widely used in academia and is also well integrated in scholarly systems such as university libraries. This database has limited obstacles in searching full text documents. Especially older, non-digitized documents are only searched via the title and the abstract. Scirus poses a similar problem, CNKI.NET did not have any detectable issues.

5.2.3 CNKI.NET

CNKI.NET is Chinas largest online database of scholar articles with a total of over 47 million articles (as of October 2013). This extensive archive strives to provide knowledge sharing and academic exchange. The platform has partnered with sites from Germany,

Hong Kong, Taiwan, Korea, Japan, and the USA. The search engine presents the most complex and extensive of all three publication databases. This complexity of the search engine itself, however, made the data acquisition very comfortable (e.g. listing of number of paper per year, etc.).

5.2.4 Scirus

Scirus is a database of more than 60 million English publications worldwide (as of October 2013). Scirus is a metasearch engine that includes search results from ScienceDirect, PubMed, Springer, Digital Archives, ArXiv.org, American Physical Society, Wiley-Blackwell, and others. This database provides the largest access to journal papers. Unfortunately the database will discontinue in early 2014.

Scirus's database has papers from 1900 right up until papers already submitted and due for publication in 2014.

5.3 Keyword Selection and Preprocess-Data Mining

As this research is concerned with the propagation of knowledge, suitable keywords were selected from a series of science and engineering books by searching their index and comparing occurrences with each other (Alpaydin, 2004; Box et al., 1978; Khine and Saleh, 2011; Myatt, 2006; Newman and Newman, 2001; Pham, 2006; Quinn, 2002; Reinhold Decker, 2007; Witten and Frank, 2011; Yip and Rubia, 2010). As knowledge is under-

stood to be scientific methodologies, algorithms, methods—generally complex knowledge that a scientist or engineer has learned and is able to apply within his analyses—standard methodologies from various fields, well-established technologies, as well as some upcoming methodologies were selected (see next subsection 5.3). The sample is extensive but not exhaustive and should give a good overview of different aspects of knowledge and technologies.

The first process of compounding the data set is stripping the text from all unimportant code fragments, extracting relevant textual data and assigning it to each variable. The new data was then written in new, better readable XML files for further processing.

Using the new data set, a cleaning process was done to remove remaining empty data points and code fragments.

The resulting data set is used to extract textual keywords—a set of keywords that was selected from prominent textbooks of science, engineering, data mining, and experimental data analysis. The instances of each keyword in every paper each year is then counted and compounded into a set of new variables. These variables are the propagation raw data.

The search for keywords was done collecting the occurrences in the abstracts of published papers and counting the total papers per year that published a paper mentioning the keyword.

The selection of the keywords themselves was done by picking the most used algorithms, research method, or analysis tool that was most present in textbooks of statistical data analysis, Soft Computing, and engineering. In the following paragraph, the keywords that have been used for this paper are listed and their contexts regarding this research are briefly explained.

Keywords Representing Knowledge This next paragraph lists all of the originally prepared keywords in alphabetical order. Refer to the glossary for brief description of each of the keywords that have been suitable for the data sets.

AdaBoost, Akaike Information Criterion, Alcubierre drive (warp drive), Analysis of Covariance and ANCOVA, ANOVA and Analysis of Variance, Application Specific Integrated Circuit, Application Specific Standard Products, Artificial Brain, Artificial Intelligence, Augmented Reality, Backpropagation, Bayes Theorem, Bayesian Information Criterion, Bayesian Network, Brain Computer Interface, C4.5, C5.0, Canonical Correlation Analysis, χ^2-Test, Complex Programmable Logic Device, Decision Theory, Decision Tree, Document Clustering, Evolutionary Algorithm, FPGA, Fractal, Fuzzy Logic, Genetic Algorithm, GPGPU, Hardware Description Language, Hebbian Learning, Hidden Markov Model, HTML, HTML 4.01, HTML 4, HTML5, Hypertext Markup Language, k-means clustering, Kansei Engineering, Large Scale Integration, Light Emitting Diode, Linear Discriminant Analysis, Machine Learning, MANCOVA, MANOVA, Markov Chains, Markov Decision Process, Medium Scale Integration, Memristor, Minimum Description Length, Monte Carlo Method, Multidimensional Scaling, Multivariate Analysis Of Covariance, MANCOVA, MySQL, Natural Language Processing, Neural Network, Nomograms, Optical Computing, Organic Light Emitting Diode, OLED, Parallel Computing, Pattern Recognition, Perceptron, Principal Component Analysis, Quantum Cryptography, Radio Frequency Identification, RFID, Self Organizing Map, Semantic Web, Small Scale Integration, Smart Grid, Social Network, Spintronics, Stem Cell Research, Support Vector Machine, Swarm Robotic, System on a Chip, Systems Integration, Ultra Large Scale Integration, Uncommitted Logic Array, Very Large Scale Integration, Virtual Reality, XHTML 1.1, XHTML 1, and XML.

During analysis using the IEEE Xplore Digital Library, the term *Fuzzy Logic* was differentiated from technologies including Fuzzy Theory. The latter was done using a compound of *Fuzzy Theory*, *Fuzzy Set*, *Fuzzy Reasoning*, and *Fuzzy Logic*.

The IEEE Xplore Digital Library did not yield enough data points for HTML and *Hypertext Markup Language* to distinguish the written one and the abbreviation, Scirus' data set was able to make a separation.

5.3.1 English, Japanese, and Chinese Keywords

The five methodologies from Soft Computing are topics directly related to–Fuzzy Logic, Neural Networks, Bayesian Network, Evolutionary Algorithm, and Genetic Algorithm. In the case of English publications, the keywords were as follows (plurals are implied):

- fuzzy theory **OR** fuzzy reasoning **OR** fuzzy logic **OR** fuzzy set
- neural network
- bayesian network **OR** bayes network **OR** belief network **OR** bayesian model **OR** bayes model
- gene expression programming **OR** genetic algorithm **OR** genetic programming **OR** evolutionary programming **OR** evolution strategy **OR** memetic algorithm **OR** differential evolution **OR** neuroevolution **OR** learning classifier system
- genetic algorithm **OR** genetic programming

The Japanese keyword compound is as follows (in the same order as the list above):

- fuzzy logic **OR** fuzzy-logic **OR** fuzzy theory **OR** fuzzy reasoning **OR** ファジー理論 **OR** ファジー推論 **OR** ファジィ理論 **OR** ファジィ推論 **OR** ファジィ論理 **OR** ファジー論理 **OR** ファジ論理 **OR** ファジ理論　ファジ推論
- neural network **OR** ニューラルネットワーク
- bayesian network **OR** bayes network **OR** belief network **OR** bayesian model **OR** bayes model **OR** ベイジアンネットワーク **OR** ベイズネットワーク
- gene expression programming **OR** genetic algorithm **OR** genetic programming **OR** evolutionary programming **OR** evolution strategy **OR** memetic algorithm **OR** differential evolution **OR** neuroevolution **OR** learning classifier system **OR** 進化アルゴリズム
- genetic algorithm **OR** genetic programming **OR** 遺伝的アルゴリズム

The Chinese set of keywords is as follows (in the same order as the list above):

- 模糊逻辑 **OR** 模糊推理 **OR** 模糊理论
- 神经网络
- 贝叶斯网络
- 遗传算法 **OR** 进化算法
- 遗传算法

The comparison of different data sets poses some obvious problems such as the structure itself, different accurateness of the representation of publications in relationship to

researcher communities, search engine algorithms, etc. To minimize the differences, the search engine algorithms were tested before setting the keywords (some search engines ignored the hyphen between words, etc.).

Language poses another difficulty. The keywords from Soft Computing were selected to provide an unambiguous set of keywords. In the case of Fuzzy Logic in Japanese, however, an uncommon complication arises. The word is translated into the Japanese alphabet "Katakana" used for foreign languages. The correct technical term for "fuzzy" is "ファジィ," but there are non-technical terms, as well as common misspellings ("ファジー," "ファジィー," etc.).

Based on extensive searches, the Chinese publication database did include the English counterparts automatically. This resulted in a short keyword list for the Chinese data set.

CHAPTER 6

RESULTS AND DISCUSSION

This section lists the results that the compartmental deterministic models have given for the data sets described in section 5. Issues with using the unchanged generic model from epidemiology in knowledge propagation domains are discussed. The generic $SEIR$ model provided good accuracy and basis for tracking knowledge in scientific publications (Marutschke and Murao, 2013). The simple approach of using the classic model also shows that with further modification, the tracking abilities can be improved. The results are shown in section 6.4.

The clustering of keywords yielded interesting results and might open up interesting prospects for other researchers. The results in section 6.2 regarding these clusters indicate that knowledge extraction is quite possible with a larger data set.

Parameter Boundaries The deterministic models with their sets of differential equations unfold their dynamic by adjusting the parameters (usually transition parameters with different implications and initial states at $t = 0$). Parameter boundaries can be necessary to ensure the correct propagation implications and dynamics. Salomon et al. (2002) have done an extensive review of cases to gather reasonable boundaries for parameters (Hepatitis C, Hepatitis B). Empirical boundary assumptions were used in Perelson et al. (1996), Chowell et al. (2003), and Bettencourt et al. (2006).

Taking previous works into consideration, together with statistics from OECD (http://www.oecdobserver.org/news/archivestory.php/aid/1160/Scientists_and_engineers.html) and the National Center for Science and Engineering Statistics (Science and Engineering Degrees from 1966-97 (http://www.nsf.gov/statistics/nsf00310/pdf/sectb.pdf)), the following boundaries were used for the model fits:

Table (6.1) Keywords representing knowledge (keyword) with propagation timespan and respective fitting performance

Parameter	Parameter Range
β	$[5, 300]$
ε	$[0, 2]$
μ	$[0, 2]$
γ	$[0, 2]$
δ	$[0, 5000]$
ϑ	$[0, 5]$

Continued on next page

Table 6.1 – continued from previous page

Parameter	Parameter Range
$S_{(t=0)}$	$[0, 10 \cdot 10^6]$
$E_{(t=0)}$	$[0, 5]$
$I_{(t=0)}$	$[0, 20]$
$R_{(t=0)}$	$[0, 5]$

Parameters δ and ϑ were empirically adjusted. Transition implications are similar for δ and β, as well as for ϑ and ε, μ, and γ. Some of the parameters had to be manually limited due to the Mathematica internal fitting algorithm being unable to provide accurate results. Fitting related issues and possible solutions are discussed in the next paragraph.

Model Fitting Mathematica-intern algorithms (NDSolve[1] and ParametricNDSolve[2]) were used for solving the sets of differential equations. The latter was used in the latest analysis due to improved computational time since the upgrade to Mathematica 9.

The algorithm for fitting differential equations to the data (as described in an earlier section) was done with NonlinearModelFit. Mathematica chooses between several fit-

[1] According to Mathematica documentation: "For ordinary differential equations, NDSolve by default uses an LSODA approach, switching between a non-stiff Adams method and a stiff Gear backward differentiation formula method." (http://reference.wolfram.com/mathematica/tutorial/SomeNotesOnInternalImplementation.html)

[2] According to Mathematica documentation: "Possible solution stages are the same as for NDSolve[...]" (http://reference.wolfram.com/mathematica/ref/ParametricNDSolve.html)

ting algorithms (amongst them "ConjugateGradient," "Gradient," "LevenbergMarquardt," "Newton," "NMinimize," and "QuasiNewton"). Mathematica's documentation does not specify on how these algorithms are selected. Several were manually tested and discarded as not suitable (Newton and Quasi Newton) or too computationally expensive (NMinimize). The *Levenberg-Marquardt* method is commonly used for solving non-linear least squares problems and showed good computation-time and fitting performance.

6.1 General $SEIR$ Model Performance

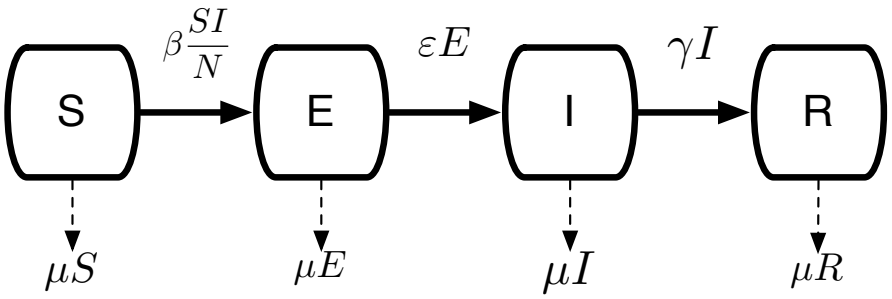


Figure (6.1) Basic $SEIR$ model with transition parameters.

Using the generic epidemiological $SEIR$ model and extracting keywords from paper abstracts enabled the tracking of knowledge in scientific publications (compartmental graphic in Figure 6.10). Having an archive of 73 years was appropriate considering that the spread of knowledge is a process that unfolds over time-periods of many decades. The IEEE Xplore Digital Library was the first database the author used to explore the intended approach and has proven to be suitable for observing and modeling knowledge

diffusion. The initial judgement was done by comparing the data set at hand with previous research (Bettencourt et al., 2006; Gurley and Johnson, 2011) and graphical assessment (data points vs. timespan). As the data set consists of publication details that were hand-selected and with the focus on information science and engineering, the propagation of certain keywords is much more pronounced than others. Some knowledge also shows signs of an endemic state, less of epidemic diffusion. In this paper, however, the authors are concerned with the knowledge that shows epidemic behavior.

Here are the equations again as in Equations (3.1a-3.1d):

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \mu N - \mu S - \beta \frac{SI}{N} \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \beta \frac{SI}{N} - (\varepsilon + \mu) E \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \varepsilon E - (\gamma + \mu) I \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \mu R
\end{aligned}$$

To calculate the Basic Reproductive Rate R_0, two transition matrices need to be calculated. Using next generation operators, we get these two matrices with new infections given by

$$\mathbf{F} = \begin{bmatrix} \beta & 0 \\ 0 & 0 \end{bmatrix}, \tag{6.1}$$

and the transfer of the infections from one compartment to another as

$$\mathbf{V} = \begin{bmatrix} 0 & \varepsilon + \mu \\ \gamma + \mu & -\varepsilon \end{bmatrix}. \tag{6.2}$$

The generation matrix is given by $\mathbf{G} = \mathbf{F}\mathbf{V}^{-1}$ and R_0 is given by the dominant Eigenvalue of $\mathbf{G}$. For the generic $SEIR$ model without birth parameter, the basic reproduction

number is given as

$$R_0 = \frac{\beta\varepsilon}{(\varepsilon+\mu)(\gamma+\mu)}. \tag{6.3}$$

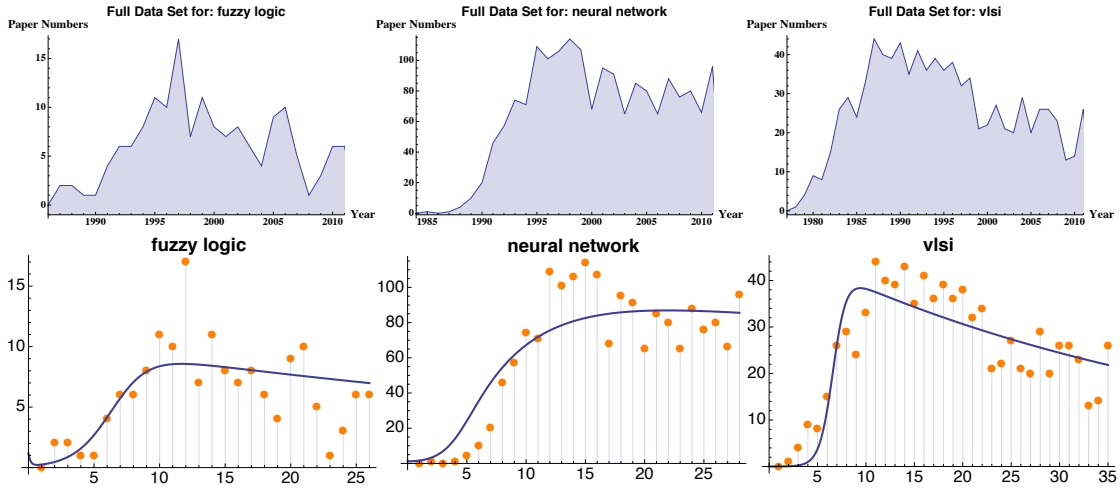


Figure (6.2) Examples of raw data (top row) and fitted I compartment of the classic *SEIR* model (bottom row).

The "Infectious" compartment as in Equation (3.1c) was successfully fitted to 17 of keywords listed in Table 6.2 with an adjusted $\bar{R}^2$ of greater than 0.75 (examples in Figure 6.2). The adjusted coefficient of determination $\bar{R}^2$ value was chosen to compensate for the number of fitting parameters and is derived from the following equations

$$R^2 = 1 - (RSS/TSS), \tag{6.4}$$

$$TSS = \sum (y_i - \bar{y})^2, \tag{6.5}$$

$$RSS = \sum (y_i - f_i)^2, \tag{6.6}$$

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{(n-1)}{(n-p-1)}, \tag{6.7}$$

with TSS being the total sum of squares, RSS the residual sum of squares, $\bar{y}$ the mean of raw data, y_i the data points of the raw data, f_i the data points of the estimation model, n the sample size, and p as the number of parameters for fitting the model to the data.

Mathematica's internal fitting algorithm with NonlinearModelFit[3] was used to fit the I-compartment to the data.

The data set being in an exploratory state, an adjusted $\bar{R}^2$ value of greater than 0.75 is considered quite appropriate for this research. The number of keywords that show a value of greater than 0.8 is 12, and 5 of the keywords have an adjusted $\bar{R}^2$ value greater than 0.9.

- Keywords with an adjusted $\bar{R}^2$ value of greater than 0.9: Neural Network, Genetic Algorithm, Principal Component Analysis, SVM, VLSI.

- Additional keywords with an adjusted $\bar{R}^2$ value of greater than 0.8: Fuzzy Logic,

[3]Mathematica chooses between several fitting algorithms (amongst them "ConjugateGradient," "Gradient," "LevenbergMarquardt," "Newton," "NMinimize," and "QuasiNewton"). Mathematica's documentation does not specify how these algorithms are selected. Several were manually tested and discarded as not suitable (Newton and Quasi Newton) or too computationally expensive (NMinimize). The *Levenberg-Marquardt* method showed good computation-time and fitting performance.

Bayesian Network, Machine Learning, Perceptron, Backpropagation, Decision Tree, SoC, FPGA,

- Additional keywords with an adjusted $\bar{R}^2$ value of greater than 0.75: Evolutionary Algorithm, Hidden Markov Model, HDL.

Table 6.2 provides an overview of all the keywords with the corresponding degree of variance and timespan of propagation.

Table (6.2) Keywords representing knowledge (keyword) with propagation timespan and respective fitting performance

Keyword	**Adjusted $\bar{R}^2$**	**Timespan propagation (years)**
Fuzzy Logic	0.83	26
Neural Network	0.97	28
Evolutionary Algorithm	0.77	18
Bayesian Network	0.88	19
Genetic Algorithm	0.97	22
AdaBoost	0.29	11
Akaike Information Criterion (AIC)	0.52	40
XML	0.64	14
HTML	0.33	17
ANOVA	0.48	15
HMM	0.29	26

Continued on next page

Table 6.2 – continued from previous page

Keyword	Adjusted $\bar{R}^2$	Timespan propagation (years)
Principal Component Analysis	0.91	26
Self Organizing Map (SOM)	0.25	17
SVM	0.93	16
Bayesian Information Criterion	0.52	22
C4.5	0.17	16
χ^2-Test	0.66	41
Document Clustering	0.44	12
Hypertext Markup Language	0.09	24
k-means	0.70	27
Machine Learning	0.89	25
Hidden Markov Model	0.78	27
MDL	0.61	25
Perceptron	0.86	21
SQL	0.51	25
Backpropagation	0.80	25
Decision Tree	0.81	47
Hebbian Learning	0.36	22
Linear Discriminant Analysis	0.56	44
Markov Decision Process	0.48	39

Continued on next page

Table 6.2 – continued from previous page

Keyword	Adjusted $\bar{R}^2$	Timespan propagation (years)
Multidimensional Scaling	0.46	41
Canonical Correlation Analysis	0.52	28
VLSI	0.97	35
SRAM	0.66	29
LSI	0.32	46
SoC	0.84	15
ASIC	0.61	15
HDL	0.76	37
FPGA	0.87	22

Not all of the keywords could be fitted with desirable performance. This highlights the potential for further data set expansions. The results shown here demonstrate that the generic $SEIR$ model is able to track the propagation of knowledge in scientific publications. Therefore, knowledge in scientific publications that is tracked via keywords from their abstracts has properties similar to that of biological epidemiology. These findings provide enough foundation to pursue the notion of knowledge diffusion using epidemiological models with the approach described in this thesis.

6.2 Clustering and Knowledge Acquisition

To explore the potential of keyword clustering, the four parameters relevant to the model (β, ε, μ, and γ) were used to cluster the knowledge related keywords. As contact rate, infectious period, incubation time, and lifetime of the infectious material have different magnitudes, the data was normalized before the categorization. Clustering was performed in sets of two, three, and four groups, using k-means clustering. The cluster using all parameters was correlated with clusters each using only one of the parameters to confirm independence. Grouping keywords using the parameters of the differential equations allows assessing the propagation characteristics of the keywords on a broad perspective.

Taking the parameters with an adjusted $\bar{R}^2$ value of greater than 0.75, a k-means clustering was performed to categorize propagation types and to associate knowledge with a group of specific propagation characteristic. Taking the four parameters—lifetime of infectious material, incubation time, recovery time, and contact rate—the k-means clustering for the normalized set of parameters was executed to divide into four, three, and two groups. Considering the number of keywords with sufficient performance, the findings for four and three-group cluster were inconclusive. Divided into two groups, however, the keywords are separated by distinctive features. Group 1 shows propagation features indicating a strong diminishing effect after passing the peak of their spread, despite their previous propagation features, speed, or timeframe. Group two shows either a continuous growth in publication numbers, or an entering into a more endemic state, i.e. the stop of growth but with no diminishing effect.

- Group 1: Fuzzy Logic, Principal Component Analysis, SVM, Hidden Markov Model,

Perceptron, Backpropagation, VLSI, SoC, FPGA

- Group 2: Neural Network, Decision Tree, HDL, Evolutionary Algorithm, Bayesian Network, Genetic Algorithm, k-means, Machine Learning

The keywords were also clustered using only one parameter independently and correlating the resulting cluster with the one using all parameters. The Pearson correlation coefficient resulted in $\beta = 0.08$, $\gamma = 0.41$, $\mu = 0.69$, $\varepsilon = 0.39$. It is not surprising that the correlation is highest with μ, the inverse rate of the number of researchers that stop publishing. Taking the rest of the correlation coefficient into account, the classification can only be explained by using the complete set of parameters.

6.3 Knowledge Propagation and Affiliated Countries

The following section describes the extraction of culture-specific attributes by applying Principal Component Analysis to a set of propagation vectors (dimensions 32 countries by 22 keywords, a timespan 1900-2013 each).

According to Lambiotte and Panzarasa (2009), scientific collaborations are vital to engage in the creation of new knowledge. The authors use a recently developed "Louvain method," publication databases and citation data to analyze network structures and how patterns of scientific collaborations contribute to knowledge creation and diffusion.

As for the selection of countries, several indicators were considered, ranging from

member countries of G-20 major economies[4], G33[5]—or a combination thereof, OECD statistics of scientifically progressed countries, SCImago Journal & Country Rank, and Thompson Reuters. As this research is based on scientific publication data, economic factors, etc. were excluded from the country selection process. The country list was formed by a cross-section of a Thompson Reuters statistic of Top 20 countries of a 10 year period in regards to *paper count*, *total citation count*, and *citation per paper*, as well as a Top 20 list of countries with the highest number of most-cited papers[6] and SCImago Journal & Country Rank ranking of countries with the most citable documents over a period from 1996-2012[7]. The selected 32 countries are listed with their statistics in the following table.

Cross referencing this list with paper numbers published in Computer Science and Engineering on scienceWATCH, five more countries were included, namely Hong Kong, Greece, Singapore, Turkey, and Portugal.

[4]The "Group of Twenty" was assumed unfit for this research due to the strong financial motivation and thus skewed sample of countries. Also the European Union as one part of the list was deemed unnecessary.

[5]The list of developing countries was deemed unrepresentative due to the strong skew of trading and economic issues, as well as little contribution to worldwide scientific publications.

[6]The data is based on two scienceWATCH web-archives (2009 and 2010). http://archive.sciencewatch.com/dr/cou/2009/09decALL/ and http://archive.sciencewatch.com/dr/cou/2010/10janALLPAPRS/

[7]Scimago Journal & Country Rank is powered by Scopus, which is presented as "the largest abstract and citation database of peer-reviewed literature" on Elsevier's homepage (http://www.scimagojr.com/countryrank.php). The countries that include the list of Thompson Reuters Top 20 lists were selected, resulting in a number of 32.

Table (6.3) Country Statistics.

Country	Paper number	Citations	Citation per Paper
Australia	276622	3067686	11.09
Austria	89782	1075042	11.97
Belgium	128800	1613458	12.53
Brazil	175063	1039235	5.94
Canada	424562	5233211	12.33
China	649689	3404466	5.24
Denmark	92734	1369297	14.77
England	682018	9399334	13.78
Finland	86509	1113141	12.87
France	548046	6304141	11.5
Germany	766162	9406841	12.28
India	253520	1288075	5.08
Ireland	57103	416966	10.52
Israel	109410	1287435	11.77
Italy	403588	4417871	10.95
Japan	788650	7602742	9.64
Korea	237652	1515555	6.38
Netherlands	236344	3419657	14.47

Continued on next page

Table 6.3 – continued from previous page

Country	Paper number	Citations	Citation per Paper
Norway	65306	764040	11.7
Poland	138705	864073	6.23
Russia	273189	1199538	4.39
Scotland	106559	1522948	14.29
Spain	305430	2942425	9.63
Sweden	174789	2407364	13.77
Switzerland	171248	2693730	15.73
Taiwan	154634	974818	6.3
USA	2974344	44669056	15.02

Principal Component Analysis (PCA) was performed on the raw data set as well as on a normalized one. The initial assessment of the Principal Components (PC) of the raw data set revealed an overly strong influence of absolute paper numbers. This reflects the first Principal Component of the raw data set with an explained variance of 99.21%. The United States dominates the number of papers with 26.1%, while the UK with the second most publications is at 8%[8].

Graphical assessment of PC2 versus PC3 of the raw data showed positive correlation of PC2 with heuristic methods such as Fuzzy Logic, Genetic Algorithm, and Neural

[8]Similar ratios can be calculated from Table 6.3 (28.7% for USA and 7.6% for second place, in this case Japan).

Network. PC2 was negatively correlated with analytical fields. Larger PC3 corresponds symbolic research fields, while smaller PC3 to more numeric research. Taking the Principal Component Scores into account, European countries such as the UK, Germany, and France are strong in symbolic and heuristic fields of the research, Asian countries place heavy focus on the field of soft-computing. Japan tends to be more similar to Brazil and Canada in terms of being compelling in analytic fields of research.

Analyzing the normalized data, the first four PCs were selected to get a cumulative variance of over 95%. Although there is no definite answer to how much variance the model should explain, a higher value (over 90%) was chosen because of the number of variables of the original data set (Quinn, 2002). From Figure 6.3, PC1 can be judged to correspond to non-analytic fields.

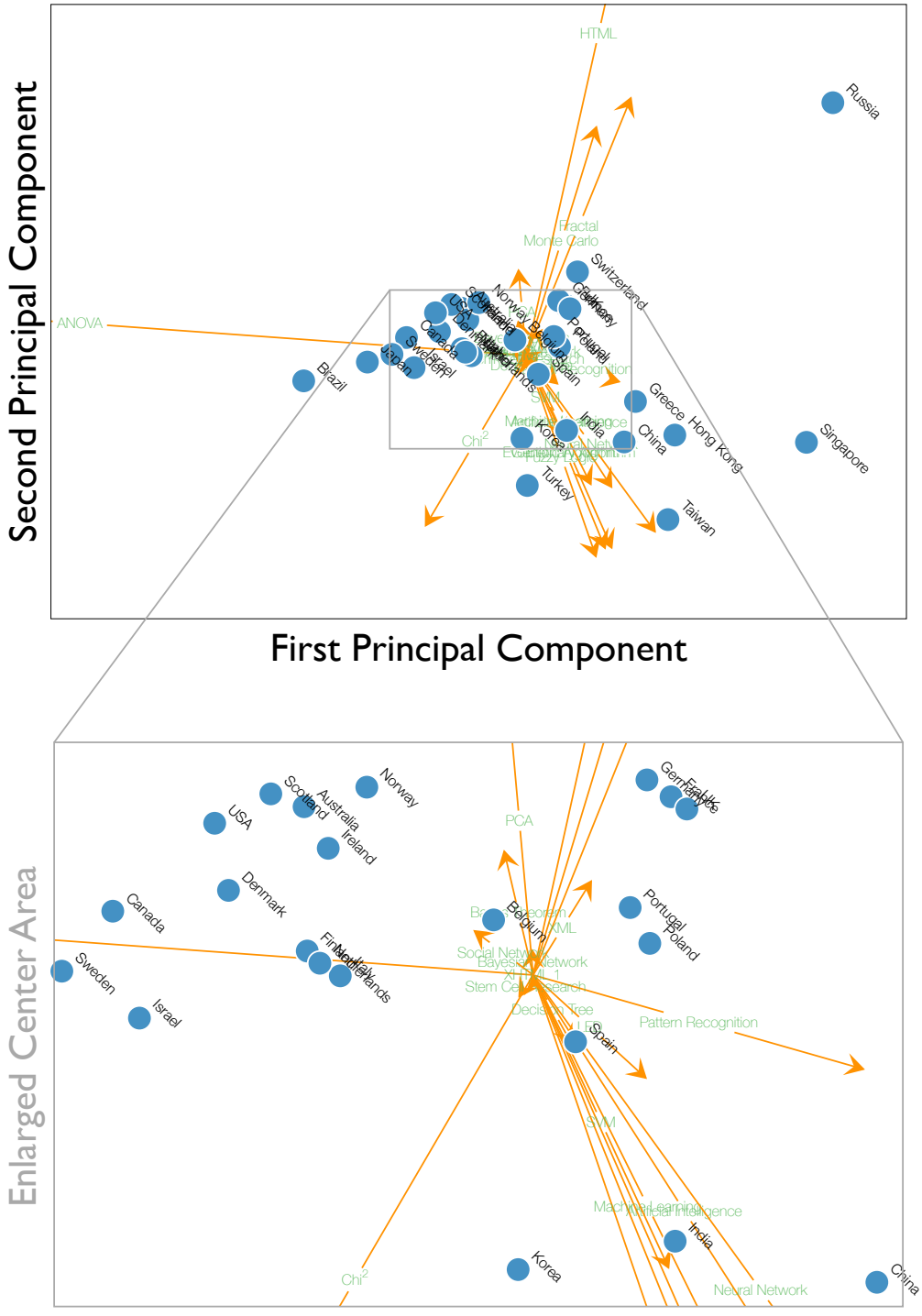


Figure (6.3) Scaled Principal Component Scores (countries marked by dots) and Loadings (fields of knowledge marked by arrows) of the first and second Principal Components.

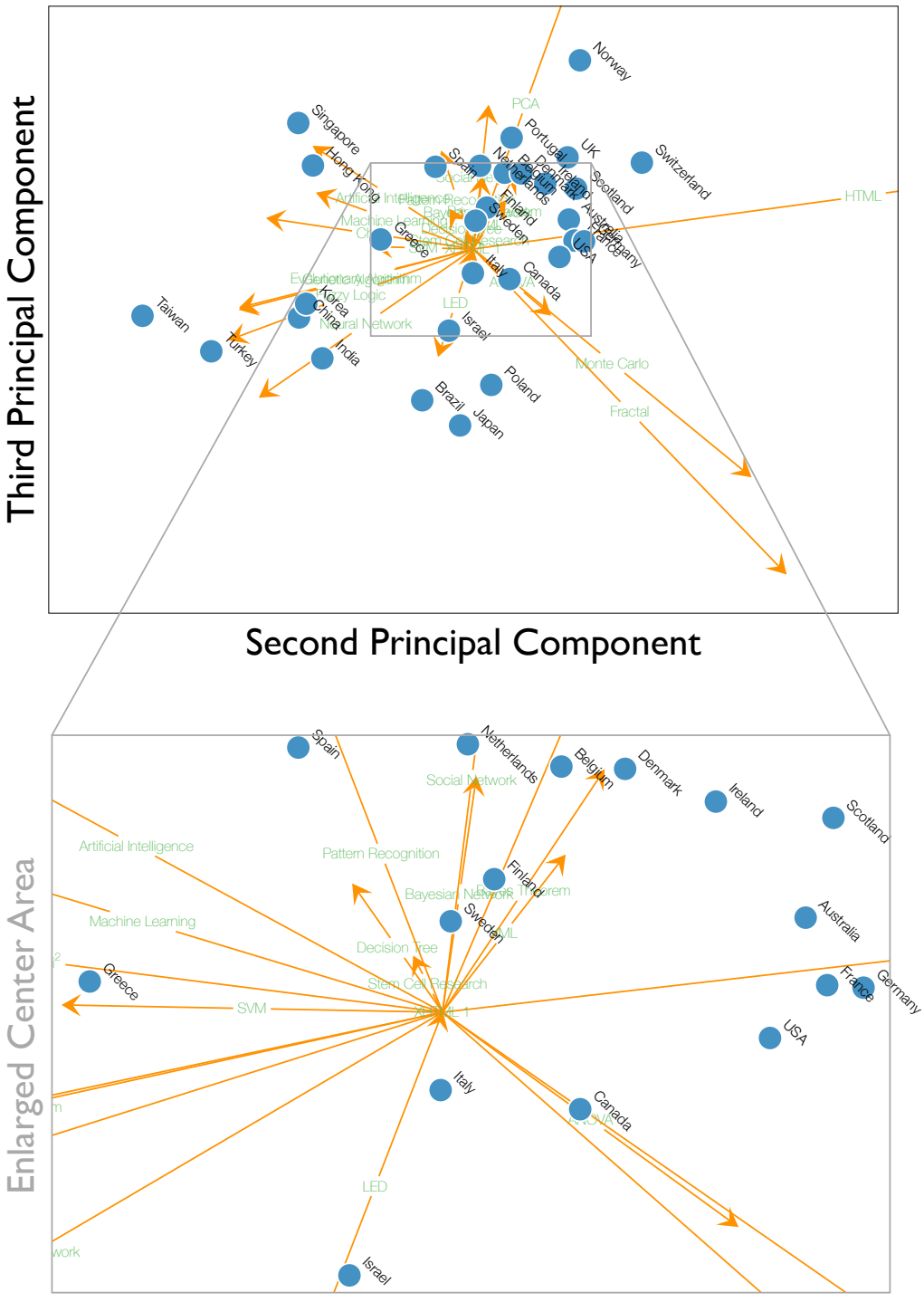


Figure (6.4) Scaled Principal Component Scores (countries marked by dots) and Loadings (fields of knowledge marked by arrows) of the second and third Principal Components.

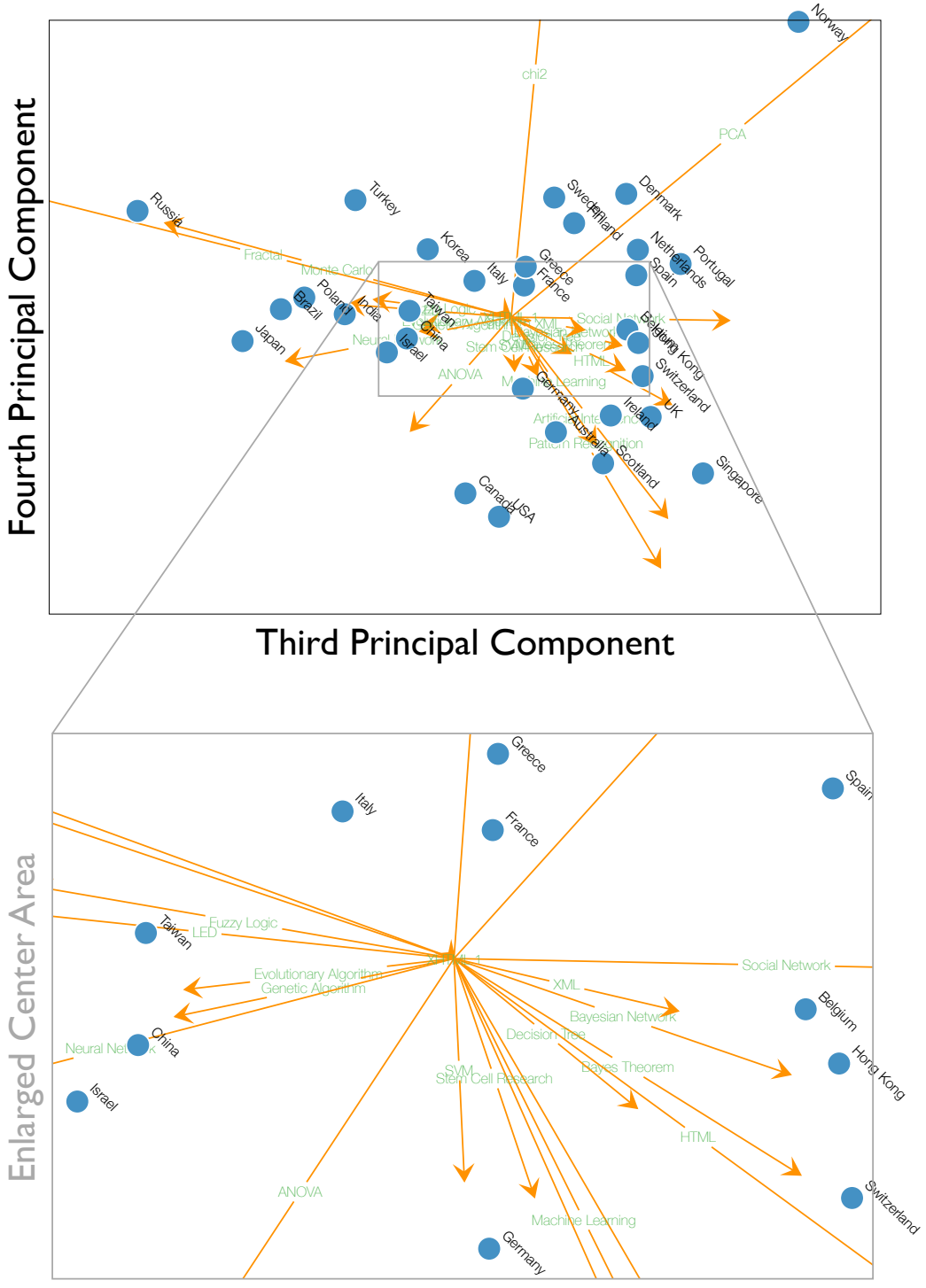


Figure (6.5) Scaled Principal Component Scores (countries marked by dots) and Loadings (fields of knowledge marked by arrows) of the third and fourth Principal Components.

6.4 *SEIR* Model compared to *SEIRE* Model

The generic *SEIR* model has shown promise in the ability to track contagion agent similar to that of complex knowledge. Some of the aspects of the tracking mechanisms used in this research, however, are not included in the causal construction of this model. This led the authors to conduct the model below.

To see the number of papers published with a specific research topic or algorithm, six unique propagation schemes were used from Soft Computing, namely the keywords "Fuzzy Logic," "Bayesian Network," "Neural Network," "Genetic Algorithm," "Evolutionary Algorithm," and a compound of keywords related to fuzzy technology.

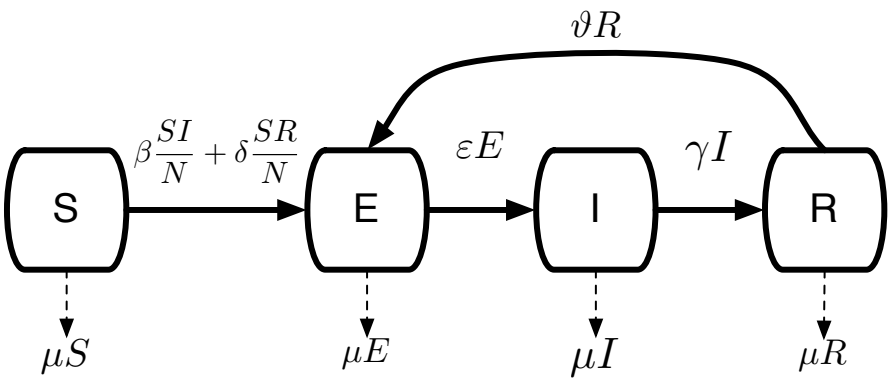
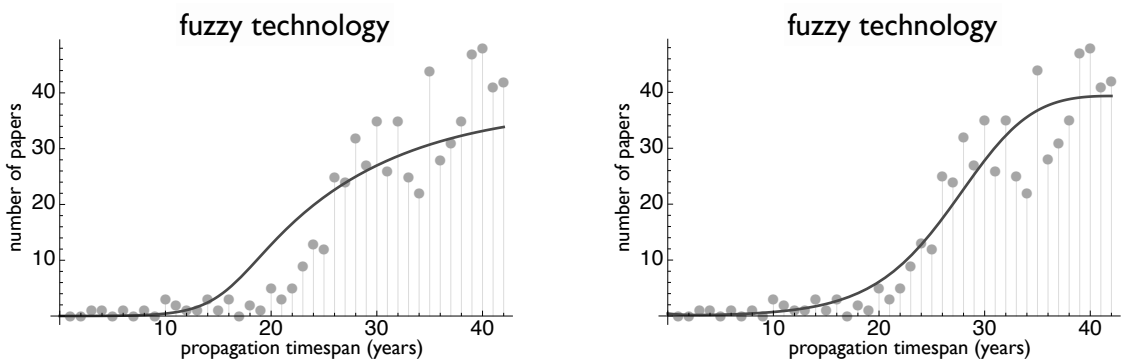


Figure (6.6) Revised *SEIR* model with transition parameters.

The graphs with the data points of the raw data and the numerically fitted model are shown in Figure 6.7 and 6.8 . All the graphs are plotted from the beginning of the propagation of the given topic (keyword) at $t = 0$, although the inception of the idea often dates back even further.



(a) $SEIR$ model. *(b)* $SEIRE$ model.

Figure (6.7) Comparison of fitting the I compartment to the propagation data of fuzzy technology.

The method of Marutschke et al. (2009) to ensure causal foundation in constructing new models was used for formulating a revised $SEIRE$ model (Figure 6.10). Epidemiological models should also follow a model of sufficient cause to include the minimal set of causations (Rothman et al., 2008). The balance is to be made between sufficient cause and unnecessary complexity. As this research tracks the number of publications regarding a research method or methodology, reflecting the knowledge propagation in that scientific community, the authors propose a loop back from the R (recovered) compartment to the E (exposed) compartment. This resembles the number of people who initially left the group of writers to propagate knowledge without publishing (additional transition potential from S to E compartment with $\delta\frac{SR}{N}$). Eventually there is the possibility that these individuals start writing a paper using the same method or methodology (transition loop back from R to E compartment with ϑR). This results in the following four differential

Equations (6.8a) to (6.8d):

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \mu S - \beta\frac{SI}{N} - \delta\frac{SR}{N} \tag{6.8a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta\frac{SI}{N} + \delta\frac{SR}{N} - (\varepsilon + \mu)E + \vartheta R \tag{6.8b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - (\gamma + \mu)I \tag{6.8c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R - \vartheta R \tag{6.8d}$$

To calculate the Basic Reproductive Rate R_0 of the *SEIRE* model, the matrix for new infections is given by

$$\mathbf{F} = \begin{bmatrix} \beta + \delta & 0 \\ 0 & 0 \end{bmatrix}, \tag{6.9}$$

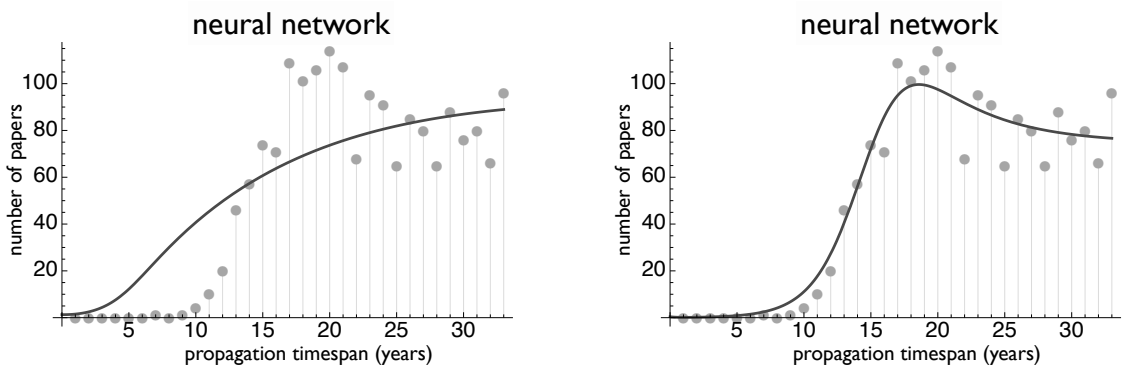
and the transfer of the infections from one compartment to another as

$$\mathbf{V} = \begin{bmatrix} 0 & \varepsilon + \mu \\ \gamma + \mu & -\varepsilon \end{bmatrix}. \tag{6.10}$$

The generation matrix is given by $\mathbf{G} = \mathbf{F}\mathbf{V}^{-1}$ and R_0 is given by the dominant Eigenvalue of $\mathbf{G}$. For the generic *SEIR* model without birth parameter, the basic reproduction number is given as

$$R_0 = \frac{\varepsilon(\beta + \delta)}{(\varepsilon + \mu)(\gamma + \mu)}. \tag{6.11}$$

The performance of the *SEIRE* model is similar to the one of the *SEIR* model when applying to the data from the initial "epidemic outbreak." In the context of the six key-



(a) $SEIR$ model. *(b)* $SEIRE$ model.

Figure (6.8) Comparison of fitting the I compartment to the propagation data of the keyword "Neural Network".

word categories, however, two features can be observed. One is in conjunction with the idea that knowledge-holders work in the background on the propagation of knowledge. Knowledge emerges at some place and needs time to reach the stage of a first published paper. This period of time is not known but demands a certain flexibility from the model that tracks the knowledge. The generic $SEIR$ model shows a weakness in this respect whereas the $SEIRE$ model tracks it with high accuracy (example of weak points of the $SEIR$ model in Figure 6.9).

The second aspect is the tracking of knowledge with a very long propagation history. Papers related to fuzzy technology have dated back over 40 years and cannot be properly tracked with the original model. The revised model is able to accurately track this particular propagation.

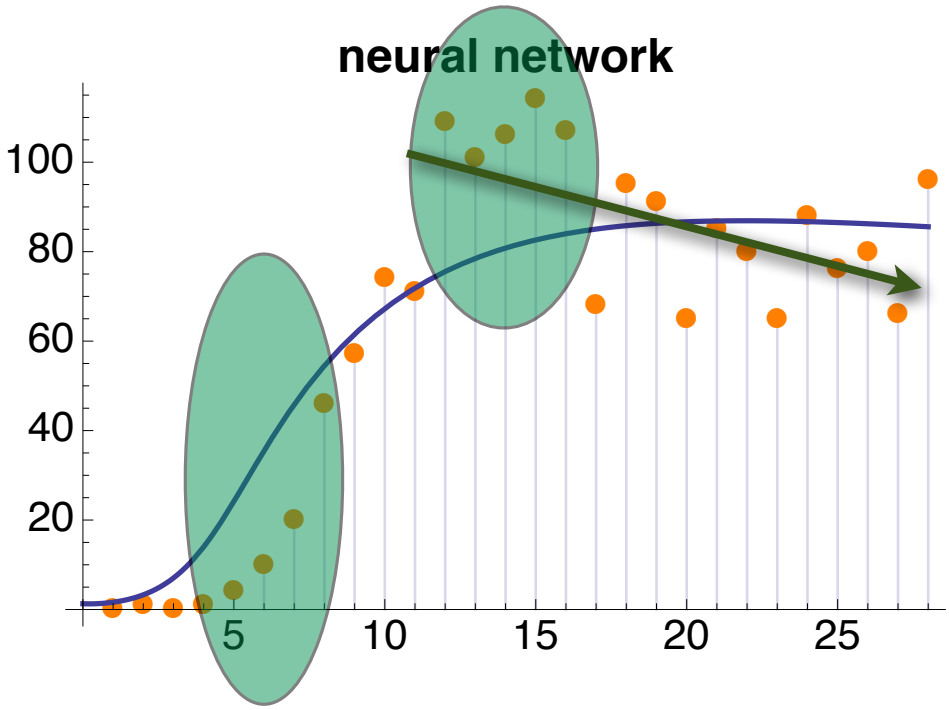
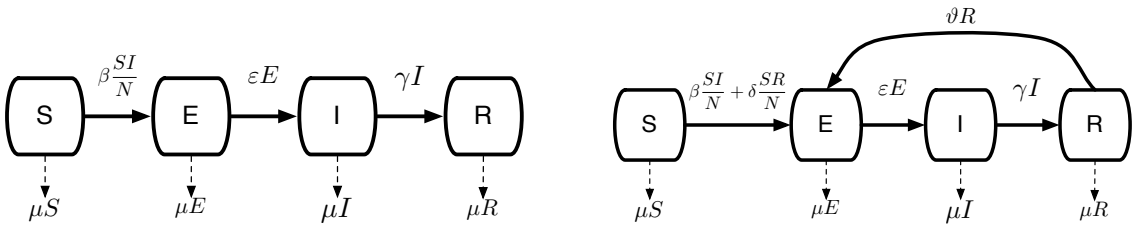


Figure (6.9) Weak areas of the classic $SEIR$ model are circled and could be improved for several keywords with the $SEIRE$ model (see Figure 6.8).

6.5 Extended Models Compared

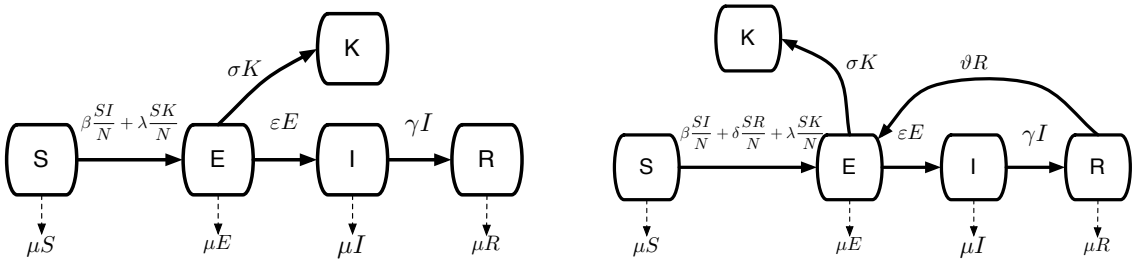
Susceptible individuals are generally able to contract a contagion agent from infected individuals and move to the next compartment resembling the incubation time. Once they can themselves transmit the contagion agent to another individual, they move to the infectious compartment. Developing antibodies and becoming immune or dying from the infection, individuals move to the removed compartment.

The differential equations that correspond with the I compartment are the ones that are fitted to the publication data. The mathematical notation is the same for all four



(a) Basic *SEIR* model. *(b)* Revised *SEIRE* model.

Figure (6.10) *SEIR*-based models (with transition parameters).



(a) Revised *SEIRK* model. *(b)* Revised *SEIREK* model.

Figure (6.11) *SEIRK*-based models (with transition parameters).

models, the dynamics unfold through the whole system.

The dynamic of this movement is formulated in four differential equations (3.1a)–(3.1d). Each equation explains the dynamic of one compartment. Outgoing arrows from each figure are denoted as negative parameters in the equations, incoming arrows with a positive sign. Equations are as in (6.8):

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \mu N - \mu S - \beta\frac{SI}{N} - \delta\frac{SR}{N} \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \beta\frac{SI}{N} + \delta\frac{SR}{N} - (\varepsilon+\mu)E + \vartheta R \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \varepsilon E - (\gamma+\mu)I \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \mu R - \vartheta R
\end{aligned}$$

The issue of including hidden knowledge-holders who can influence the spreading of information while not appearing in the infectious compartment is addressed with the *SEIRE* model as in Figure 6.10. Researchers who have published a paper moved to the R compartment but still exercise influence on the susceptible individuals ($\delta\frac{SR}{N}$). They also have the ability to go back to the paper writing process, indicated by the arrow which loops back to the E compartment with ratio ϑR (Figure 6.10).

The dynamic of this system is described by four differential equations as in equations (6.8).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \mu S - \beta\frac{SI}{N} - \lambda\frac{SK}{N} \tag{6.12a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta\frac{SI}{N} + \lambda\frac{SR}{N} - (\varepsilon+\mu)E + \sigma K \tag{6.12b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - (\gamma+\mu)I \tag{6.12c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R \tag{6.12d}$$

$$\frac{\mathrm{d}K}{\mathrm{d}t} = \sigma K - \mu K \tag{6.12e}$$

With the notion of hidden knowledge-holders who move directly from the paper writing process to a position where influence is practiced but a publication is never pursued, the *SEIRK* model was proposed. Although the performance and robustness was improved over the basic epidemiological model, the *SEIRE* model showed better overall performance (Table 6.4). This would fall into the causality of researchers again publishing a paper when they have acquired a certain research methodology. A discussion about causality is given in the next section.

Table (6.4) Keyword categories and comparison of adjusted $\bar{R}^2$ of the *SEIR*, *SEIRE*, *SEIRK*, and *SEIREK* models.

Keyword	$\bar{R}^2$ (*SEIR*)	$\bar{R}^2$ (*SEIRE*)	$\bar{R}^2$ (*SEIRK*)	$\bar{R}^2$ (*SEIREK*)
Evolutionary Algorithm	0.77	0.81	0.80	0.81
Genetic Algorithm	0.97	0.98	0.98	0.98
Fuzzy Logic	0.83	0.85	0.85	0.85
Fuzzy Technology	0.91	0.95	0.95	0.95
Neural Network	0.89	0.98	0.95	0.97
Bayesian Network	0.91	0.91	0.91	0.91

Regarding the whole population N of the *SEIR* based models (Equations (3.1) and (6.8)), equation $\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}E}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} = 0$ applies as well as $N = S + E + I + R$. Likewise for the *SEIRK* based models (Equations (6.12) and (6.13)), equation $\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}E}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} + \frac{\mathrm{d}K}{\mathrm{d}t} = 0$ applies as well as $N = S + E + I + R + K$.

Figure 6.11 shows the compartments with transition parameters, leading to equations (6.12).

Table (6.5) Soft Computing keyword categories and the corresponding adjusted $\bar{R}^2$ value in Japan, China, and worldwide.

Keyword	$\bar{R}^2$ **(Japan)**	$\bar{R}^2$ **(China)**	$\bar{R}^2$ **(Worldwide)**
Evolutionary Algorithm	0.82	0.83	0.88
Genetic Algorithm	0.89	0.83	0.92
Fuzzy Logic	0.86	0.98	0.87
Neural Network	0.83	0.97	0.95
Bayesian Network	0.84	0.88	0.86

Combining the two *SEIRE* and *SEIRK* models, both influential parties with and without the possibility for re-publishing are considered. The resulting *SEIREK* model is shown in Figure 6.11 with equations (6.13).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \mu S - \beta\frac{SI}{N} - \delta\frac{SR}{N} - \lambda\frac{SK}{N} \tag{6.13a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta\frac{SI}{N} + \delta\frac{SR}{N} + \lambda\frac{SK}{N} - (\varepsilon + \mu)E + \vartheta R \tag{6.13b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - (\gamma + \mu)I \tag{6.13c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R - \vartheta R \tag{6.13d}$$

$$\frac{\mathrm{d}K}{\mathrm{d}t} = \sigma K - \mu K \tag{6.13e}$$

The cultural analysis resulted in high performances for tracking the knowledge (Table 6.5) and keywords could be correctly assigned to the cultural area.

6.6 Analysis of $SEIR$ and $SEIRE$ model with the Scirus Data Set

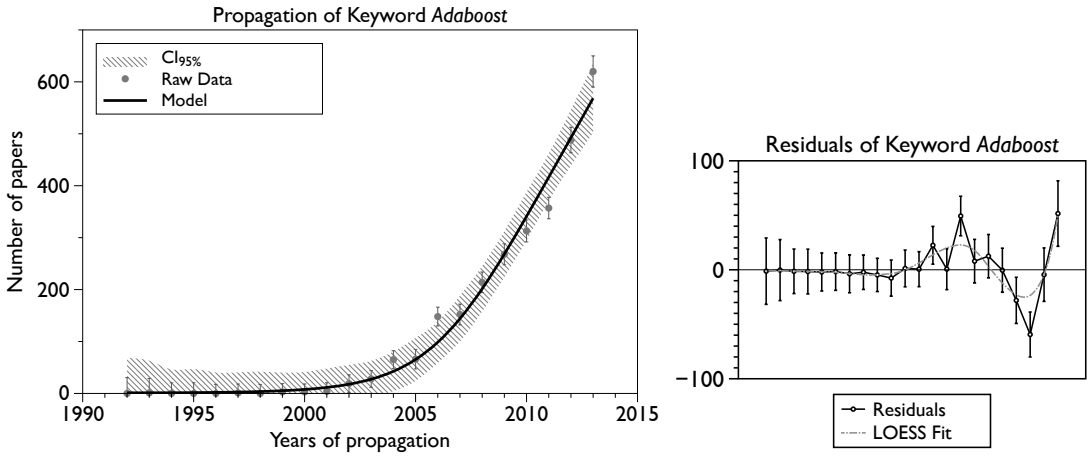
Both the $SEIR$ and $SEIRE$ model were fitted to the Scirus data set with new sets of basic reproduction number R_0 and evaluation of the residuals. As can be seen from most of the residuals plot, some unexplained variance seem to be consistent. The $SEIRE$ model showed better or equal performance most of the time, being able to track knowledge that has a slow growth in the beginning. The causality implications, however, are quite different. From the construction of the Basic Reproductive Rate R_0 (Equation 6.11) can be inferred, that the influence of individuals that hold the knowledge but do not publish a paper is quite large (expressed by ε).

Figures 6.12-6.69 show the data plots corresponding to each keyword with the $SEIRE$ and $SEIR$ model fit, 95% confidence interval (Figures 6.12a-6.69a) and the matching distribution of the residuals (Figures 6.12b-6.69b).

To emphasize trending, the residuals are traced with a LOESS fit with smoothing parameter of 0.3 and second degree local polynomials[9] .

The data from each keyword was also randomly resampled 10 times, with sample size

[9]The parameters for the LOESS fit were determined empirically to visualize residual tendency over time. Literature was consulted for standard parameter ranges (Quinn, 2002).



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.12) Data and *SEIRE* model of keyword "AdaBoost."

matching the original vector, to minimize systematic errors. Adjusted $\bar{R}^2$ values, the basic reproduction number R_0, and parameters of the model-fit were calculated and the mean of all 10 resamples were produced.

In case of the keyword "AdaBoost" and the *SEIRE* model, the $\bar{R}^2 = 0.980$ and $R_0 = 3030.04$ (Figure 6.12).

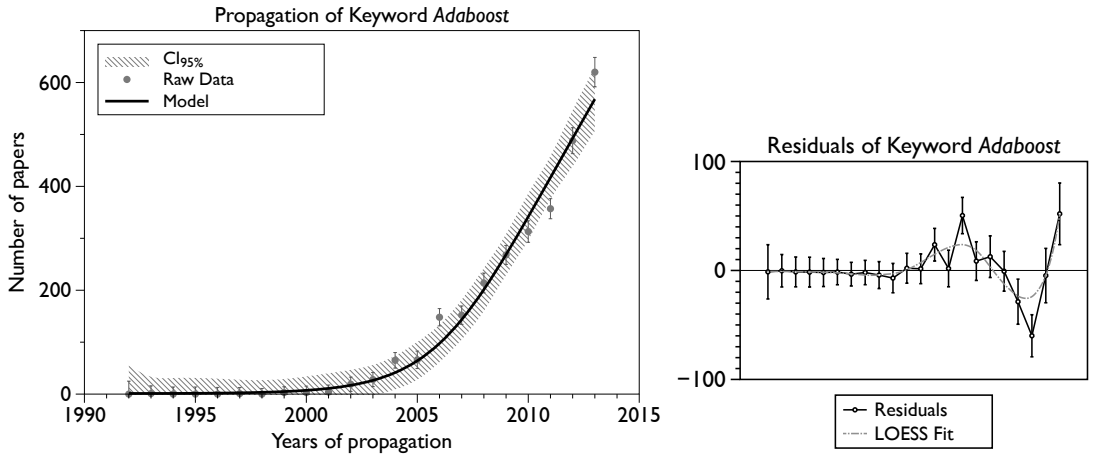
The resample produced $\bar{R}^2 = 0.974$ and $R_0 = 3458.78$.

In case of the keyword "AdaBoost" and the *SEIR* model, the $\bar{R}^2 = 0.982$ and $R_0 = 2060.16$ (Figure 6.13).

The resample produced $\bar{R}^2 = 0.987$ and $R_0 = 1926.16$.

For "AdaBoost," both models fit similarly to the data.

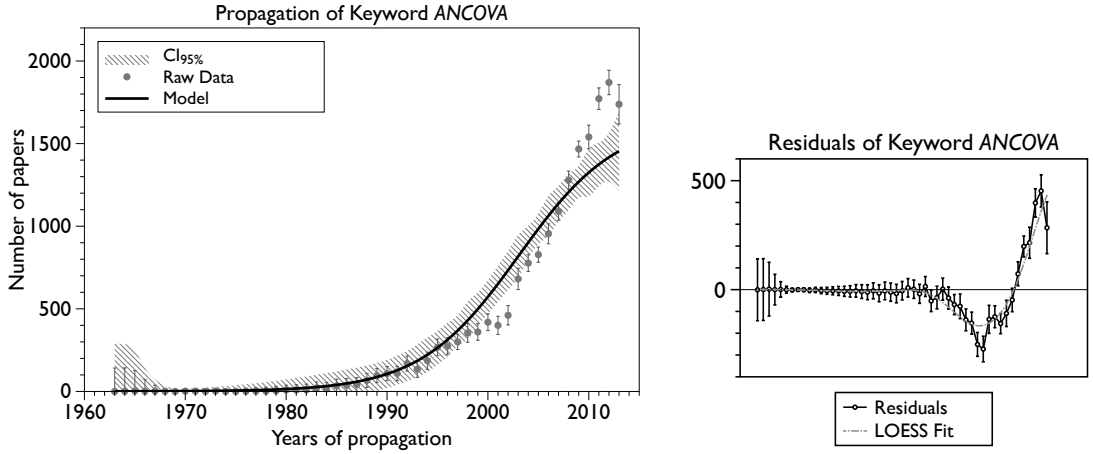
In case of the keyword "ANCOVA" and the *SEIRE* model, the $\bar{R}^2 = 0.951$ and $R_0 = 13.018$ (Figure 6.14).



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

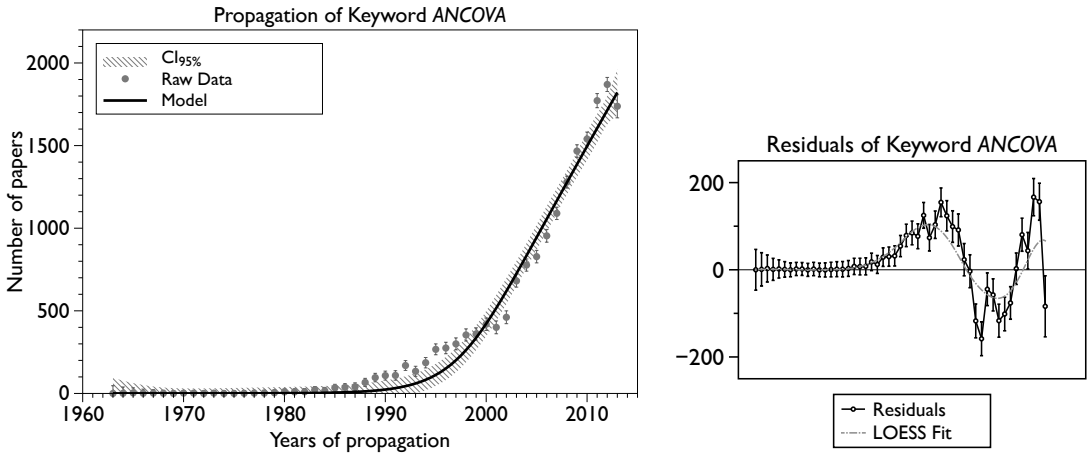
Figure (6.13) Data and *SEIR* model of keyword "AdaBoost."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.14) Data and *SEIRE* model of keyword "ANCOVA."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.15) Data and *SEIR* model of keyword "ANCOVA."

The resample produced $\bar{R}^2 = 0.946$ and $R_0 = 105.375$.

In case of the keyword "ANCOVA" and the *SEIR* model, the $\bar{R}^2 = 0.985$ and $R_0 = 2946.11$ (Figure 6.15).

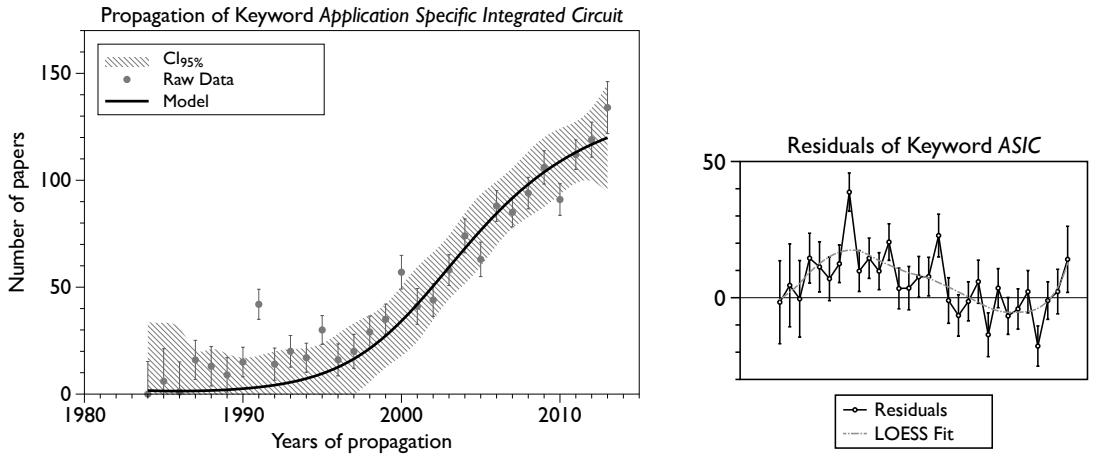
The resample produced $\bar{R}^2 = 0.979$ and $R_0 = 10184$.

Both cases of "ANCOVA" have unexplained variance, which can be seen from the residuals. The *SEIRE* model seems to overemphasize the inception, whereas the *SEIR* model fits to the later fast propagation.

In case of the keyword "Application Specific Integrated Circuit" and the *SEIRE* model, the $\bar{R}^2 = 0.939$ and $R_0 = 637.522$ (Figure 6.16).

The resample produced $\bar{R}^2 = 0.917$ and $R_0 = 508.322$.

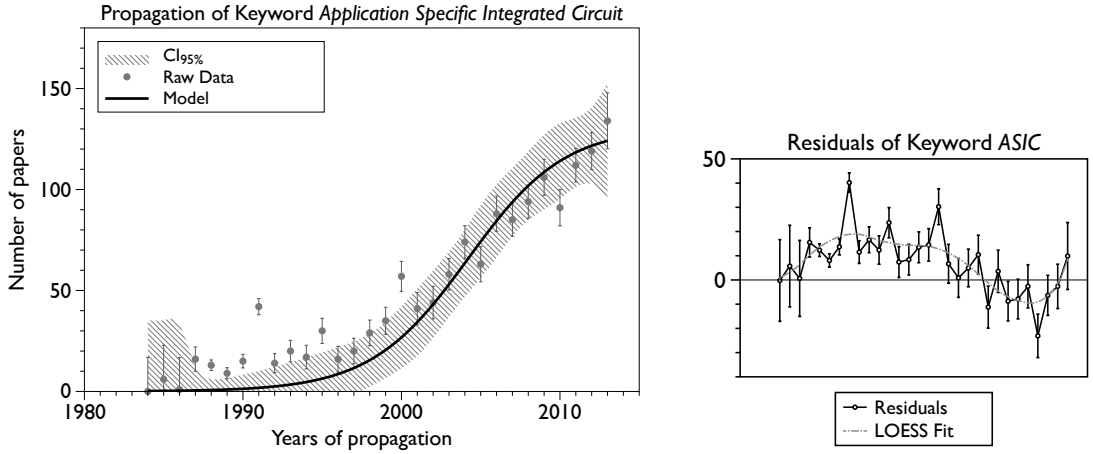
In case of the keyword "Application Specific Integrated Circuit" and the *SEIR* model, the $\bar{R}^2 = 0.925$ and $R_0 = 1.39306$ (Figure 6.17).



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

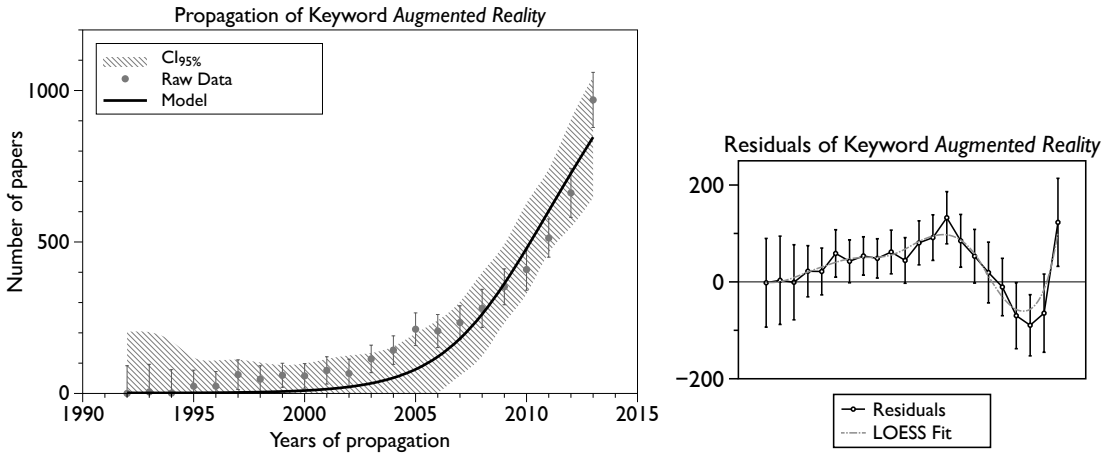
Figure (6.16) Data and *SEIRE* model of keyword "Application Specific Integrated Circuit."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.17) Data and *SEIR* model of keyword "Application Specific Integrated Circuit."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.18) Data and *SEIRE* model of keyword "Augmented Reality."

The resample produced $\bar{R}^2 = 0.911$ and $R_0 = 1.74596$.

Slightly higher degree of explained variance and slightly more randomly distributed residuals can be interpreted in favor of the *SEIRE* model. From this, the implication here is a much higher adoption rate of this particular technology.

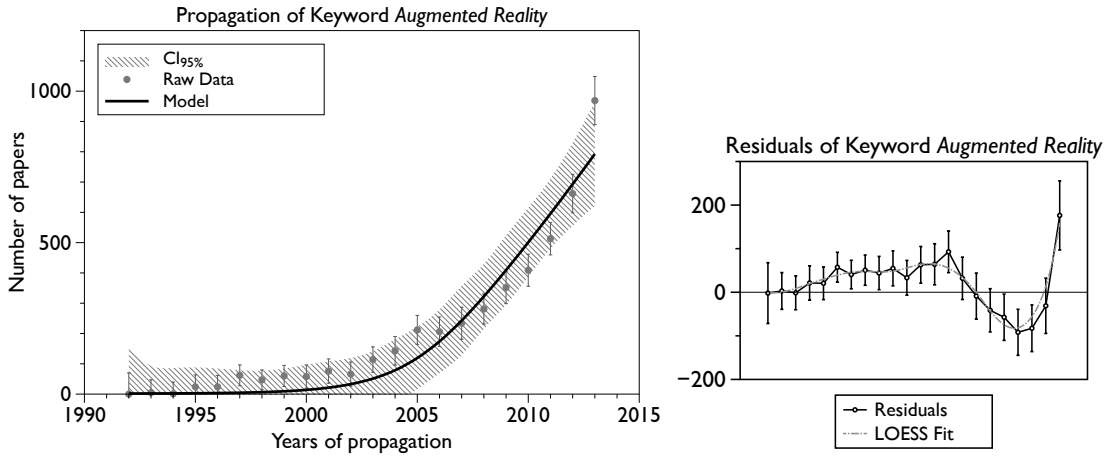
In case of the keyword "Augmented Reality" and the *SEIRE* model, the $\bar{R}^2 = 0.917$ and $R_0 = 1039.53$ (Figure 6.18).

The resample produced $\bar{R}^2 = 0.944$ and $R_0 = 10812.9$.

In case of the keyword "Augmented Reality" and the *SEIR* model, the $\bar{R}^2 = 0.935$ and $R_0 = 10412.6$ (Figure 6.19).

The resample produced $\bar{R}^2 = 0.927$ and $R_0 = 8308.52$.

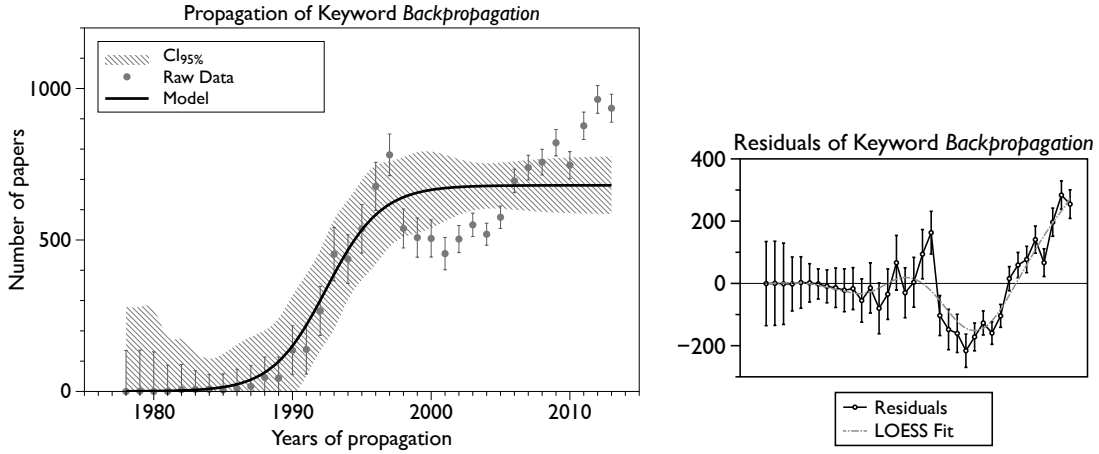
The *SEIR* model shows better performance for "Augmented Reality" and points to a high adoption rate.



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

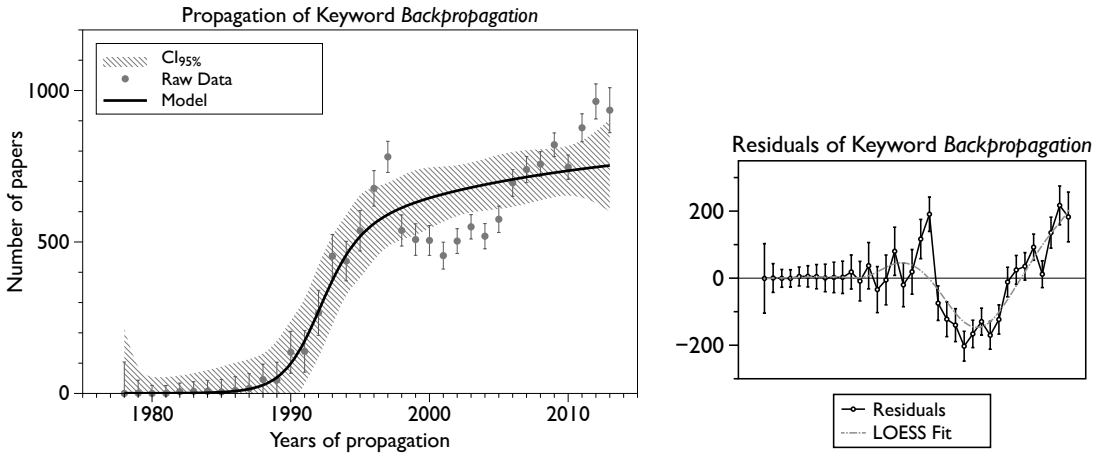
Figure (6.19) Data and *SEIR* model of keyword "Augmented Reality."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.20) Data and *SEIRE* model of keyword "Backpropagation."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.21) Data and *SEIR* model of keyword "Backpropagation."

In case of the keyword "Backpropagation" and the *SEIRE* model, the $\bar{R}^2 = 0.930$ and $R_0 = 14.8043$ (Figure 6.20).

The resample produced $\bar{R}^2 = 0.937$ and $R_0 = 3.89249$.

In case of the keyword "Backpropagation" and the *SEIR* model, the $\bar{R}^2 = 0.952$ and $R_0 = 9.30039$ (Figure 6.21).

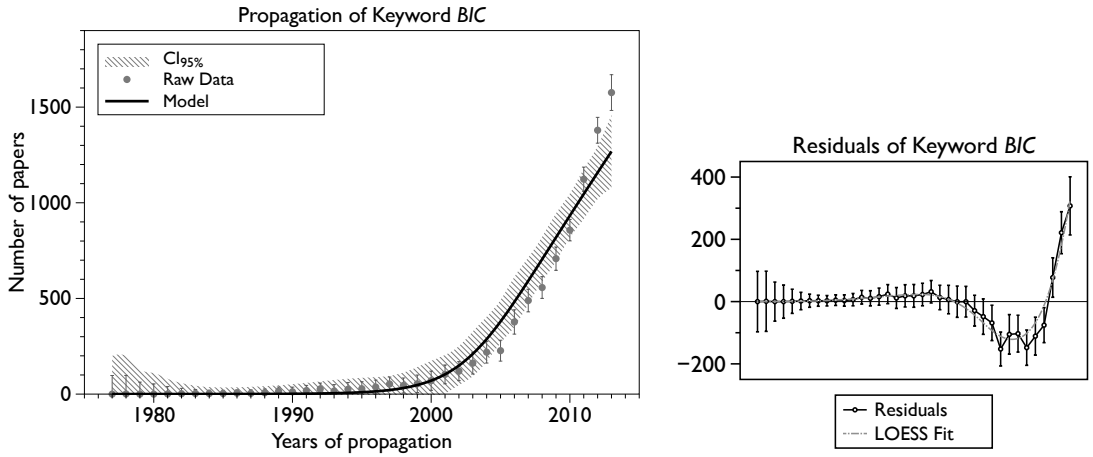
The resample produced $\bar{R}^2 = 0.937$ and $R_0 = 2.75756$.

Both models fail to track the first peak in case of "Backpropagation."

In case of the keyword "Bayesian Information Criterion" and the *SEIRE* model, the $\bar{R}^2 = 0.955$ and $R_0 = 2958.74$ (Figure 6.22).

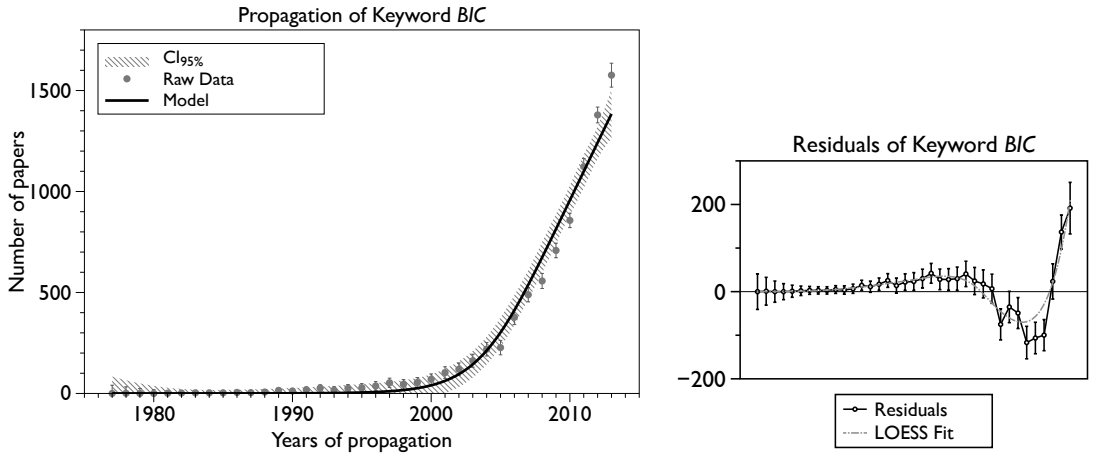
The resample produced $\bar{R}^2 = 0.954$ and $R_0 = 7925.75$.

In case of the keyword "Bayesian Information Criterion" and the *SEIR* model, the $\bar{R}^2 = 0.981$ and $R_0 = 3959.6$ (Figure 6.23).



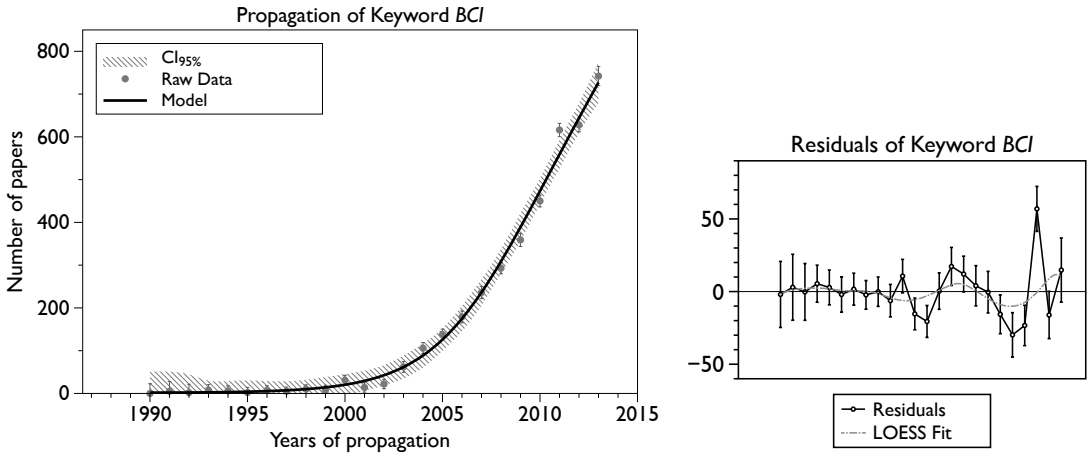
(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars *(b)* Residuals, Standard Error bars, and LOESS curve

Figure (6.22) Data and *SEIRE* model of keyword "Bayesian Information Criterion."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars *(b)* Residuals, Standard Error bars, and LOESS curve

Figure (6.23) Data and *SEIR* model of keyword "Bayesian Information Criterion."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.24) Data and *SEIRE* model of keyword "Brain Computer Interface."

The resample produced $\bar{R}^2 = 0.975$ and $R_0 = 4353.05$.

In case of the keyword "Brain Computer Interface" and the *SEIRE* model, the $\bar{R}^2 = 0.993$ and $R_0 = 2500.26$ (Figure 6.24).

The resample produced $\bar{R}^2 = 0.988$ and $R_0 = 2145.18$.

In case of the keyword "Brain Computer Interface" and the *SEIR* model, the $\bar{R}^2 = 0.993$ and $R_0 = 756.06$ (Figure 6.25).

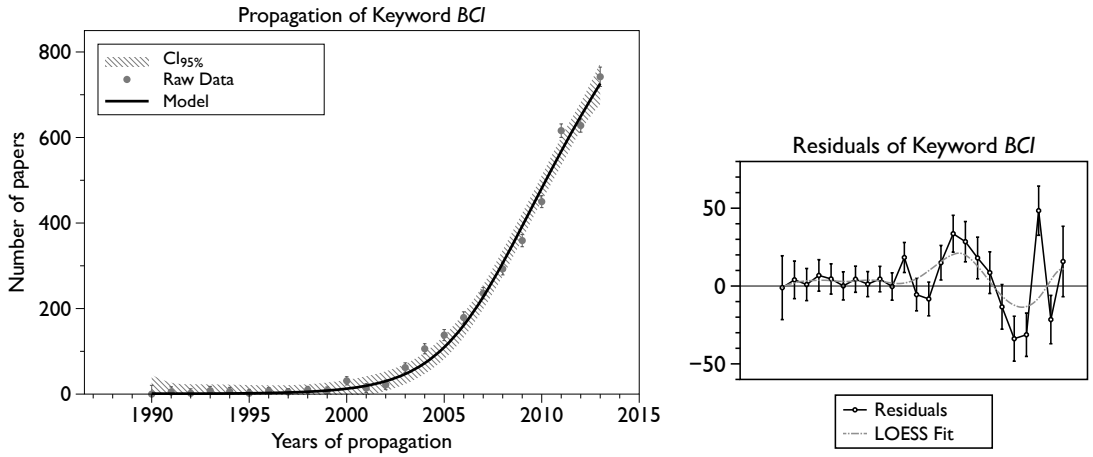
The resample produced $\bar{R}^2 = 0.995$ and $R_0 = 352.509$.

For "BCI" the residuals are more randomly distributed for the *SEIRE* model.

In case of the keyword "Canonical Correlation Analysis" and the *SEIRE* model, the $\bar{R}^2 = 0.938$ and $R_0 = 3.97869$ (Figure 6.26).

The resample produced $\bar{R}^2 = 0.887$ and $R_0 = 3.50052$.

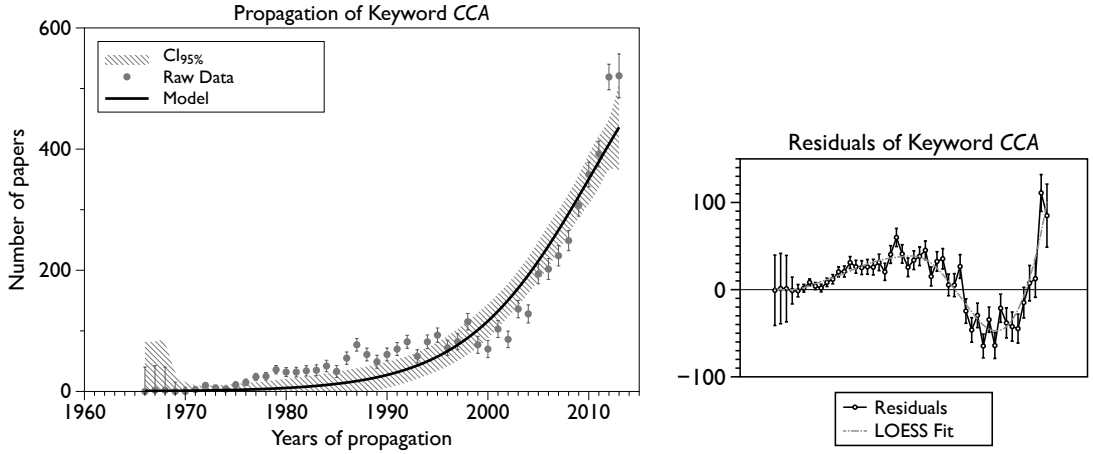
In case of the keyword "Canonical Correlation Analysis" and the *SEIR* model, the



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

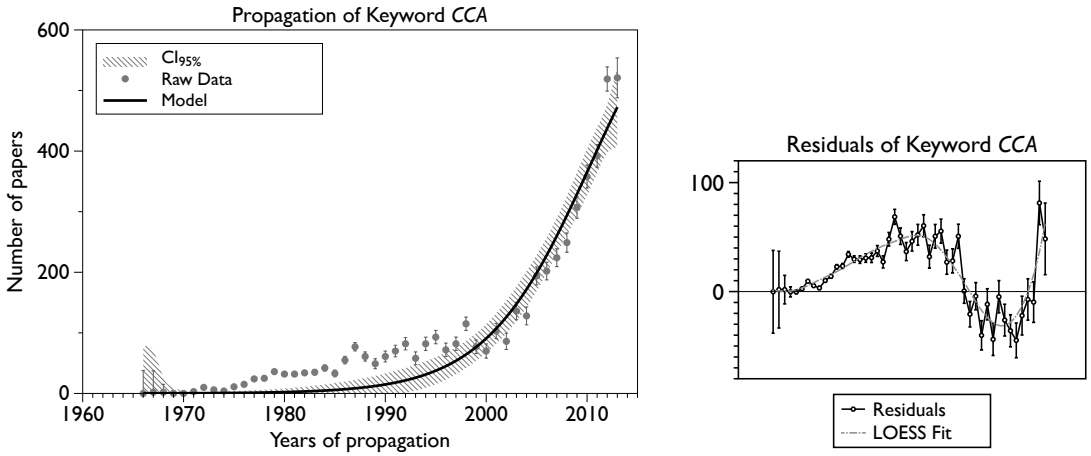
Figure (6.25) Data and *SEIR* model of keyword "Brain Computer Interface."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.26) Data and *SEIRE* model of keyword "Canonical Correlation Analysis."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.27) Data and *SEIR* model of keyword "Canonical Correlation Analysis."

$\bar{R}^2 = 0.945$ and $R_0 = 1.53887$ (Figure 6.27).

The resample produced $\bar{R}^2 = 0.943$ and $R_0 = 1.41077$.

In case of the keyword "Complex Programmable Logic Device" and the *SEIRE* model, the $\bar{R}^2 = 0.878$ and $R_0 = 5.178$ (Figure 6.28).

The resample produced $\bar{R}^2 = 0.745$ and $R_0 = 5.999$.

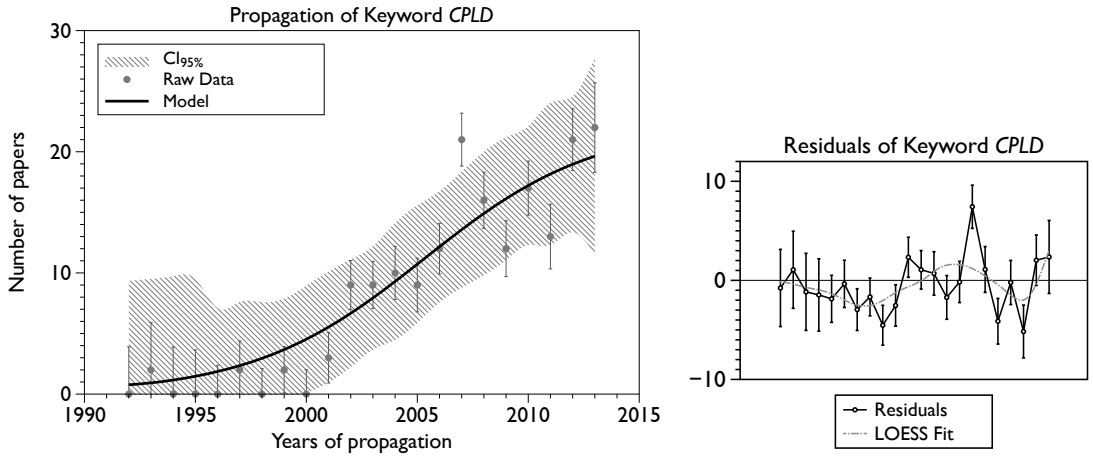
In case of the keyword "Complex Programmable Logic Device" and the *SEIR* model, the $\bar{R}^2 = 0.907$ and $R_0 = 1.48488$ (Figure 6.29).

The resample produced $\bar{R}^2 = 0.774$ and $R_0 = 1.50465$.

In case of the keyword "Document Clustering" and the *SEIRE* model, the $\bar{R}^2 = 0.978$ and $R_0 = 1.246$ (Figure 6.30).

The resample produced $\bar{R}^2 = 0.946$ and $R_0 = 0.687$.

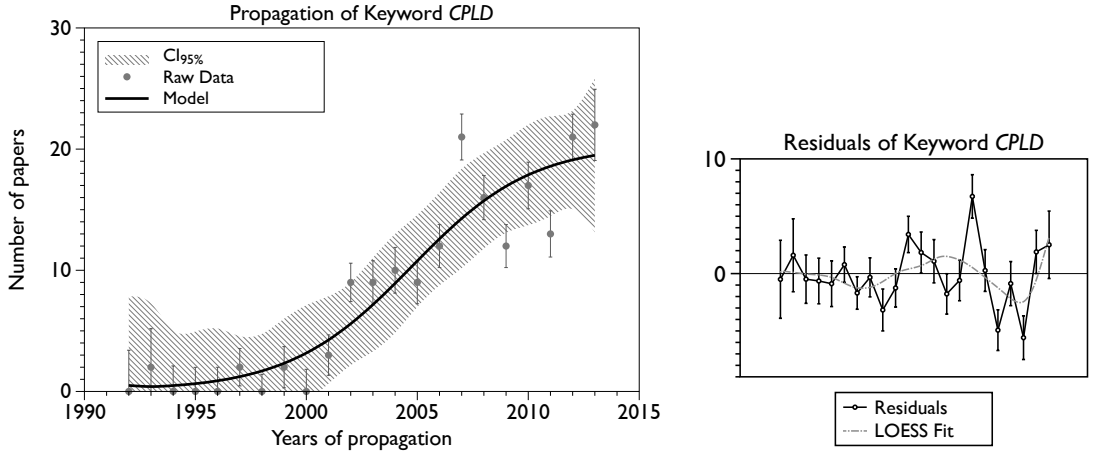
In case of the keyword "Document Clustering" and the *SEIR* model, the $\bar{R}^2 = 0.922$



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

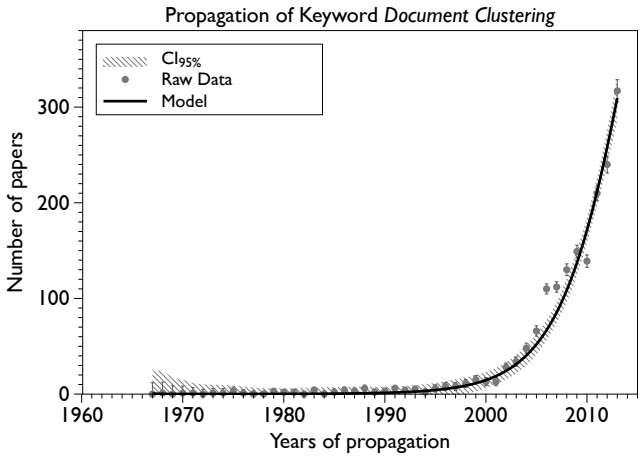
Figure (6.28) Data and *SEIRE* model of keyword "Complex Programmable Logic Device."



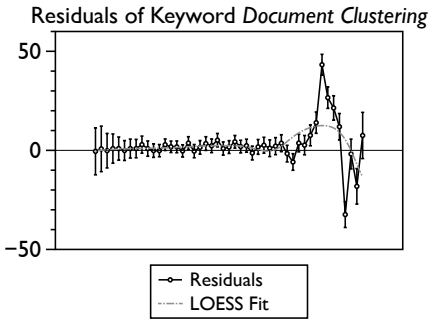
(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.29) Data and *SEIR* model of keyword "Complex Programmable Logic Device."

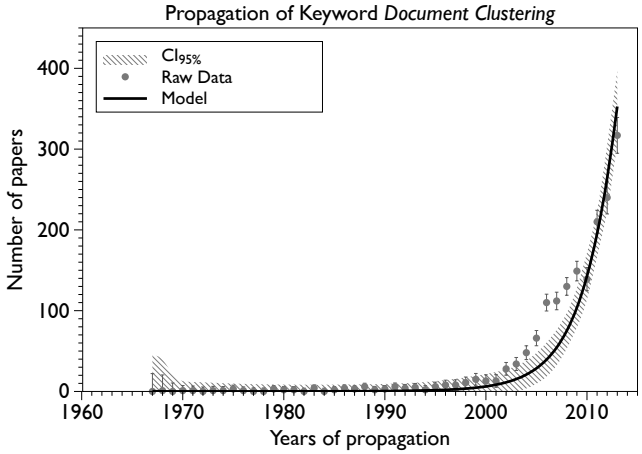


(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

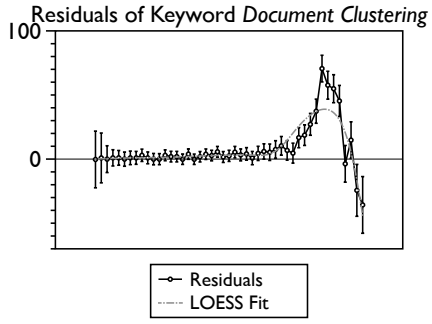


(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.30) Data and *SEIRE* model of keyword "Document Clustering."

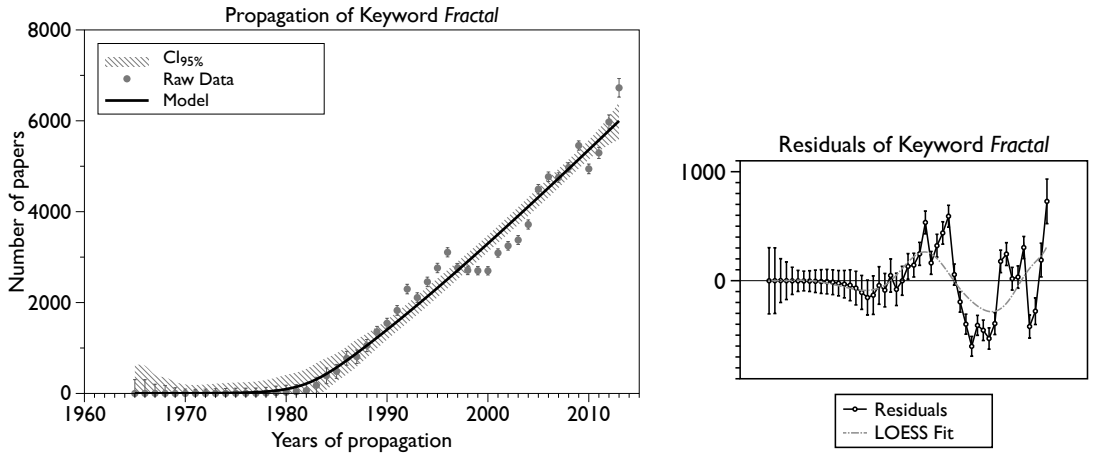


(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars



(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.31) Data and *SEIR* model of keyword "Document Clustering."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.32) Data and *SEIRE* model of keyword "Fractal."

and $R_0 = 0.970489$ (Figure 6.31).

The resample produced $\bar{R}^2 = 0.958$ and $R_0 = 1.13257$.

In case of the keyword "Fractal" and the *SEIRE* model, the $\bar{R}^2 = 0.988$ and $R_0 = 24315.600$ (Figure 6.32).

The resample produced $\bar{R}^2 = 0.987$ and $R_0 = 37590.800$.

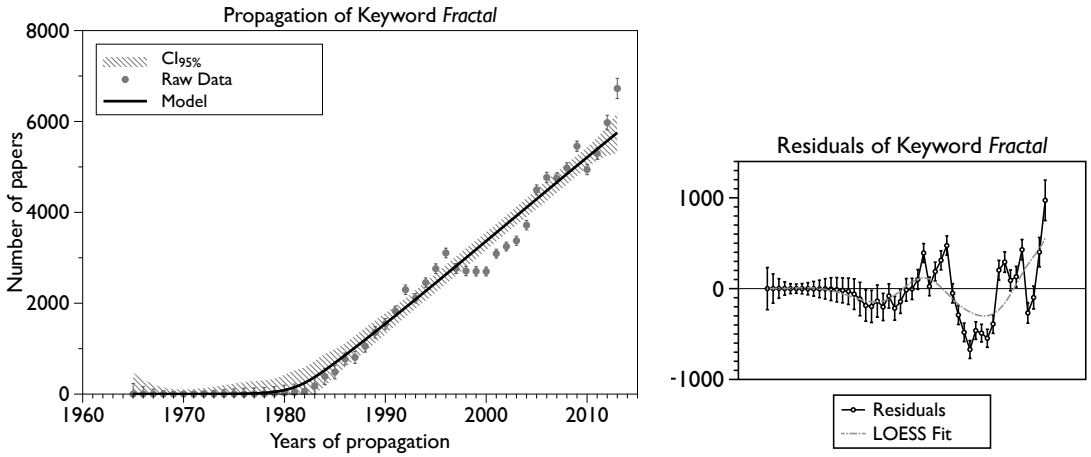
In case of the keyword "Fractal" and the *SEIR* model, the $\bar{R}^2 = 0.987$ and $R_0 = 8710.75$ (Figure 6.33).

The resample produced $\bar{R}^2 = 0.983$ and $R_0 = 8533.42$.

In case of the keyword "Fuzzy Logic" and the *SEIRE* model, the $\bar{R}^2 = 0.931$ and $R_0 = 5563.610$ (Figure 6.34).

The resample produced $\bar{R}^2 = 0.922$ and $R_0 = 1284.720$.

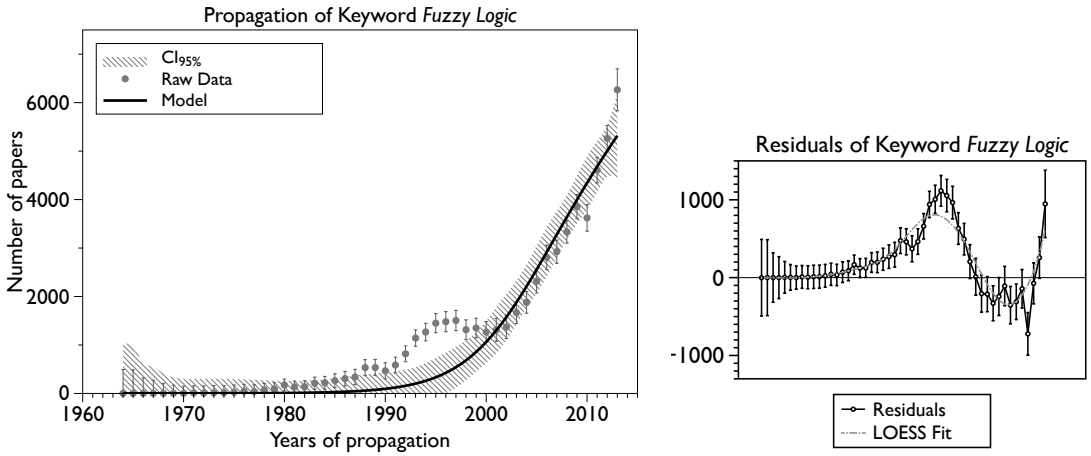
In case of the keyword "Fuzzy Logic" and the *SEIR* model, the $\bar{R}^2 = 0.934$ and



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

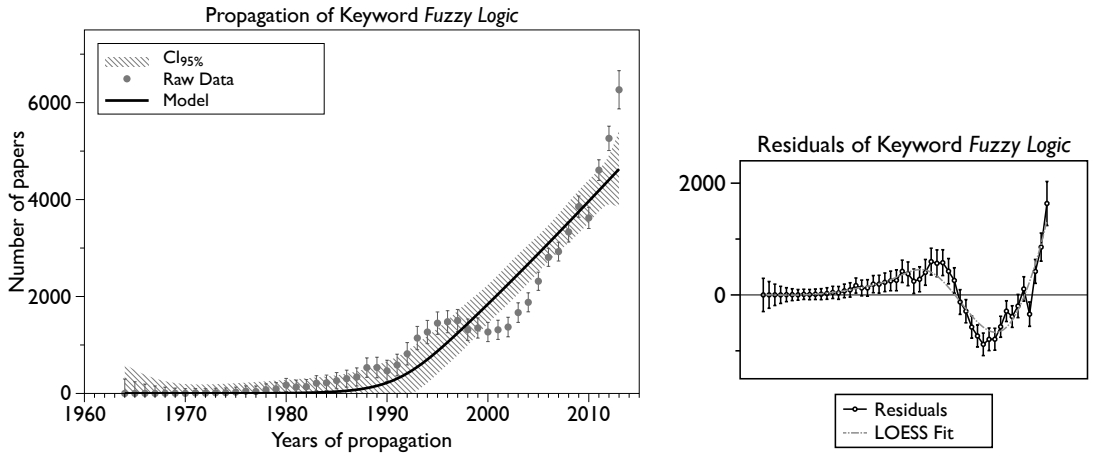
Figure (6.33) Data and *SEIR* model of keyword "Fractal."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.34) Data and *SEIRE* model of keyword "Fuzzy Logic."



(a) Data, $SEIR$ model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.35) Data and $SEIR$ model of keyword "Fuzzy Logic."

$R_0 = 9907.1$ (Figure 6.35).

The resample produced $\bar{R}^2 = 0.900$ and $R_0 = 10473.5$.

In case of the keyword "Genetic Algorithm" and the $SEIRE$ model, the $\bar{R}^2 = 0.983$ and $R_0 = 2.459$ (Figure 6.36).

The resample produced $\bar{R}^2 = 0.984$ and $R_0 = 3.375$.

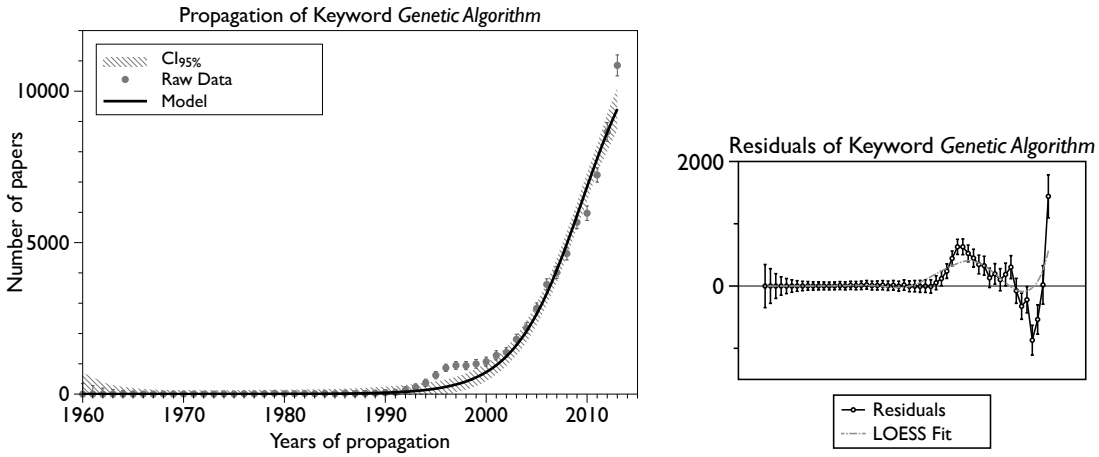
In case of the keyword "Genetic Algorithm" and the $SEIR$ model, the $\bar{R}^2 = 0.965$ and $R_0 = 1.05309$ (Figure 6.37).

The resample produced $\bar{R}^2 = 0.816$ and $R_0 = 1.33665$.

In case of the keyword "k-means Clustering" and the $SEIRE$ model, the $\bar{R}^2 = 0.994$ and $R_0 = 5.419$ (Figure 6.38).

The resample produced $\bar{R}^2 = 0.994$ and $R_0 = 13.615$.

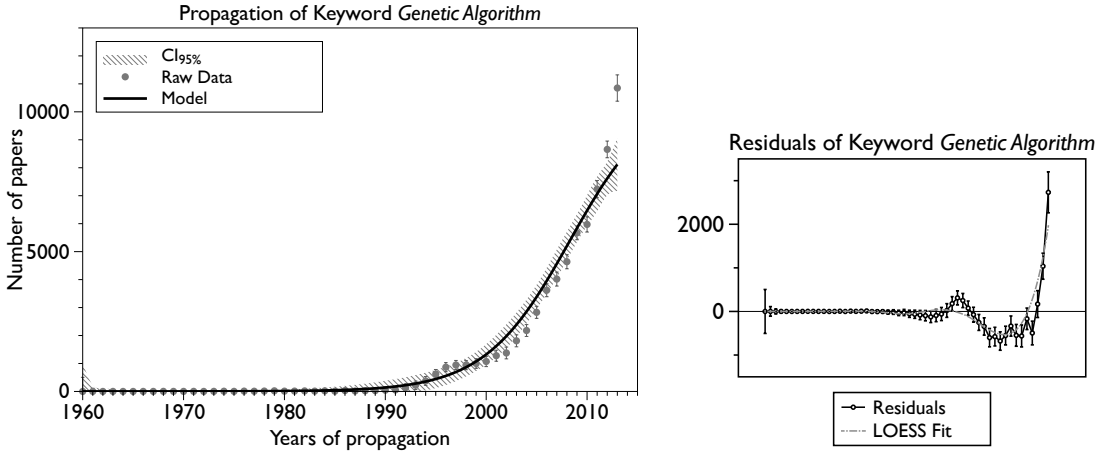
In case of the keyword "k-means Clustering" and the $SEIR$ model, the $\bar{R}^2 = 0.990$



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

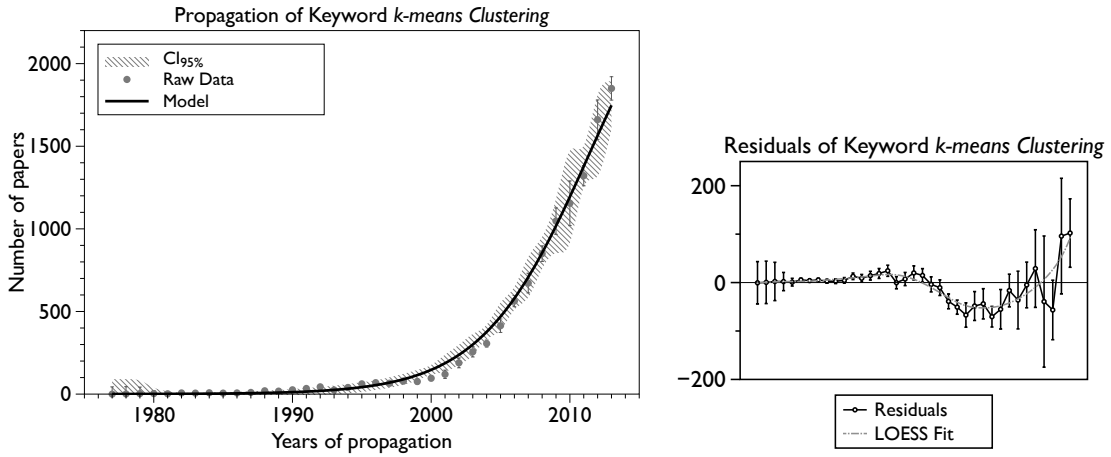
Figure (6.36) Data and *SEIRE* model of keyword "Genetic Algorithm."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

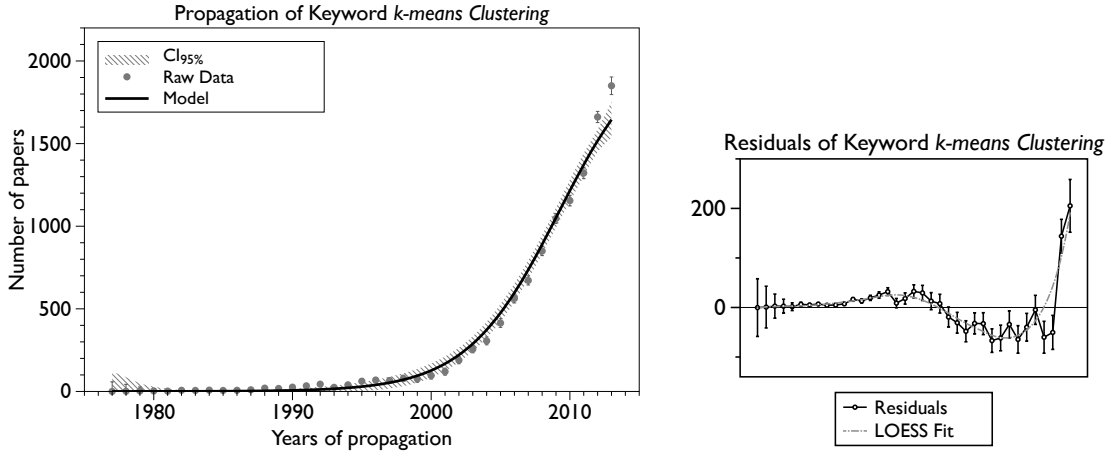
Figure (6.37) Data and *SEIR* model of keyword "Genetic Algorithm."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

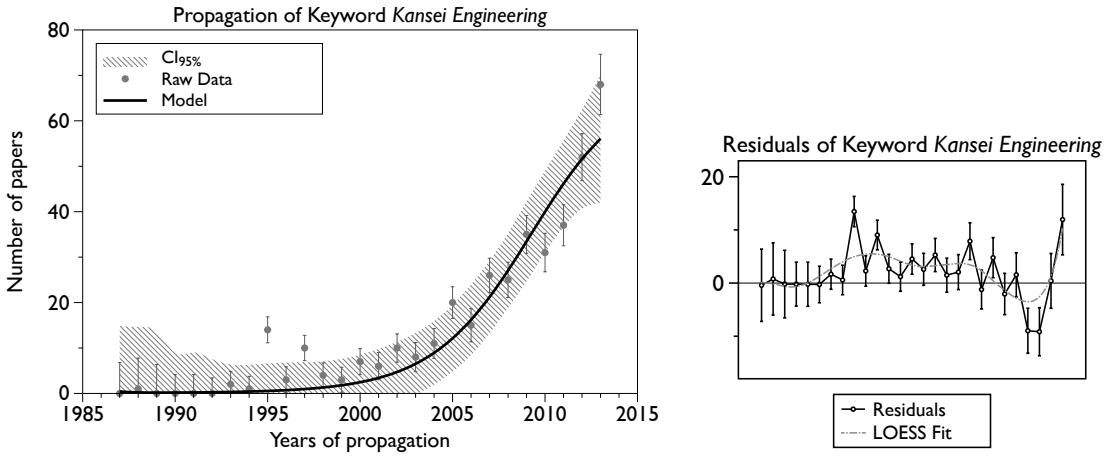
Figure (6.38) Data and *SEIRE* model of keyword "k-means Clustering."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.39) Data and *SEIR* model of keyword "k-means Clustering."



(a) Data, $SEIRE$ model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.40) Data and $SEIRE$ model of keyword "Kansei Engineering."

and $R_0 = 1.67244$ (Figure 6.39).

The resample produced $\bar{R}^2 = 0.994$ and $R_0 = 1.27641$.

In case of the keyword "Kansei Engineering" and the $SEIRE$ model, the $\bar{R}^2 = 0.907$ and $R_0 = 1.682$ (Figure 6.40).

The resample produced $\bar{R}^2 = 0.921$ and $R_0 = 8.502$.

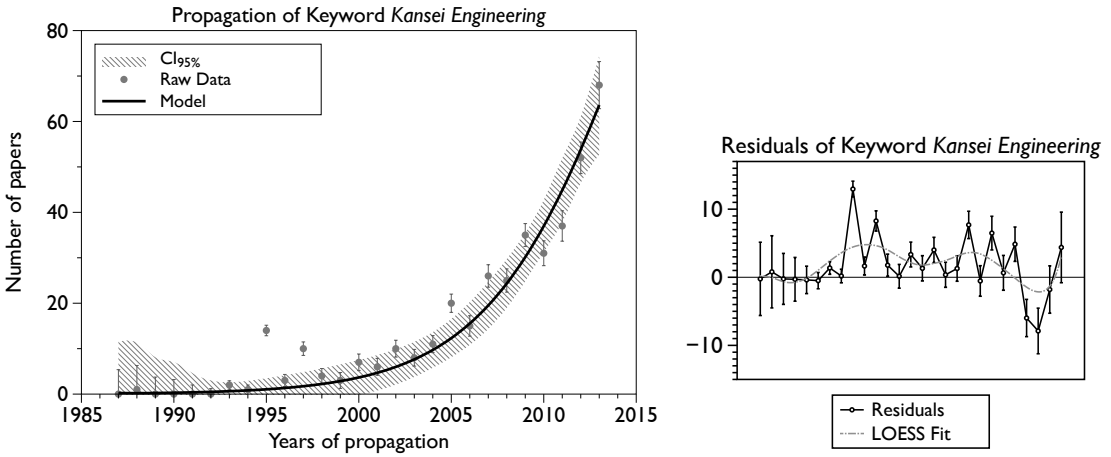
In case of the keyword "Kansei Engineering" and the $SEIR$ model, the $\bar{R}^2 = 0.942$ and $R_0 = 1.76922$ (Figure 6.41).

The resample produced $\bar{R}^2 = 0.875$ and $R_0 = 17.698$.

In case of the keyword "Light Emitting Diode" and the $SEIRE$ model, the $\bar{R}^2 = 0.989$ and $R_0 = 1.174$ (Figure 6.42).

The resample produced $\bar{R}^2 = 0.990$ and $R_0 = 0.948$.

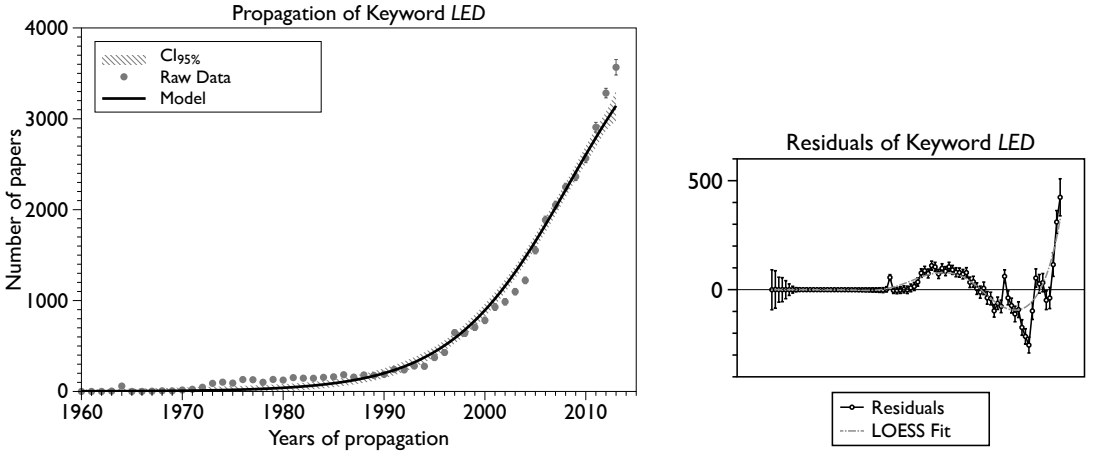
In case of the keyword "Light Emitting Diode" and the $SEIR$ model, the $\bar{R}^2 = 0.984$



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

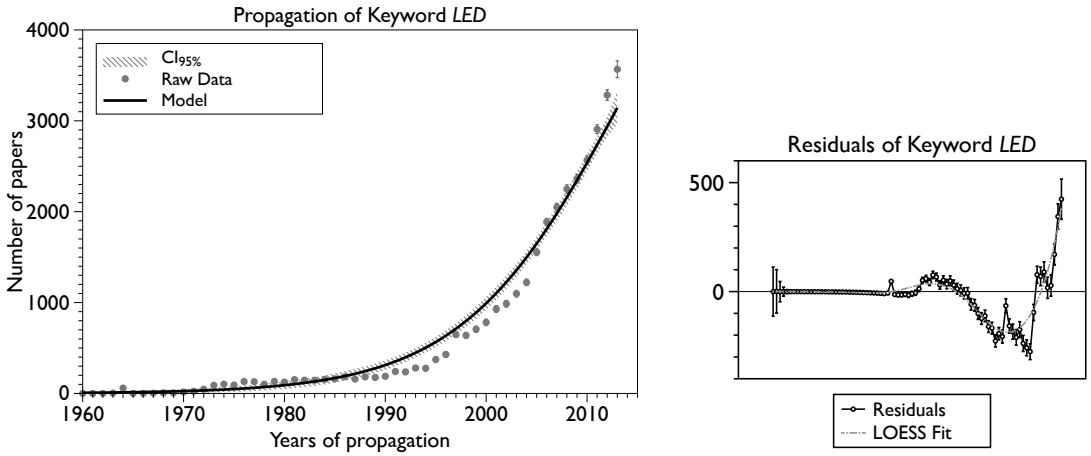
Figure (6.41) Data and *SEIR* model of keyword "Kansei Engineering."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.42) Data and *SEIRE* model of keyword "Light Emitting Diode."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.43) Data and *SEIR* model of keyword "Light Emitting Diode."

and $R_0 = 3.75703$ (Figure 6.43).

The resample produced $\bar{R}^2 = 0.966$ and $R_0 = 1.01732$.

In case of the keyword "Minimum Description Length" and the *SEIRE* model, the $\bar{R}^2 = 0.988$ and $R_0 = 3.791$ (Figure 6.44).

The resample produced $\bar{R}^2 = 0.981$ and $R_0 = 5.672$.

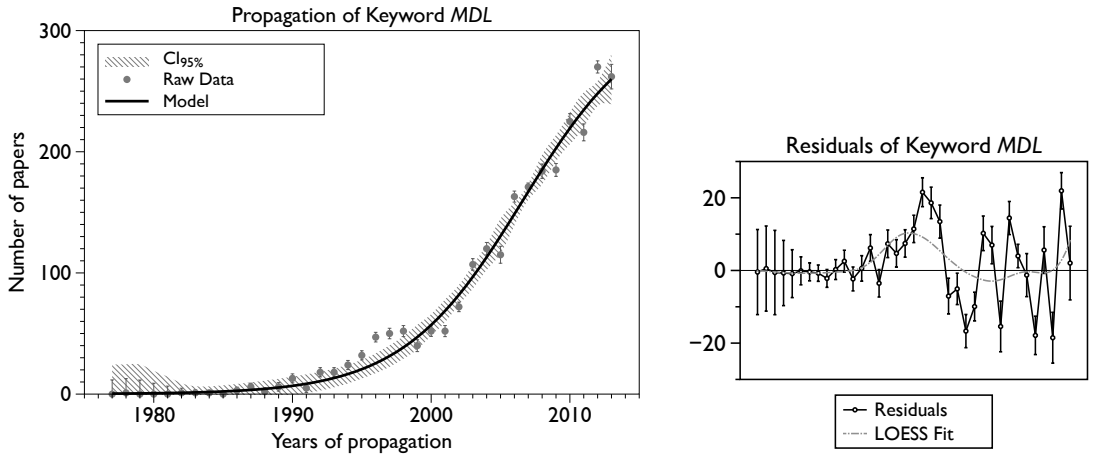
In case of the keyword "Minimum Description Length" and the *SEIR* model, the $\bar{R}^2 = 0.986$ and $R_0 = 1.43314$ (Figure 6.45).

The resample produced $\bar{R}^2 = 0.962$ and $R_0 = 1.7253$.

In case of the keyword "MANCOVA" and the *SEIRE* model, the $\bar{R}^2 = 0.973$ and $R_0 = 32.248$ (Figure 6.46).

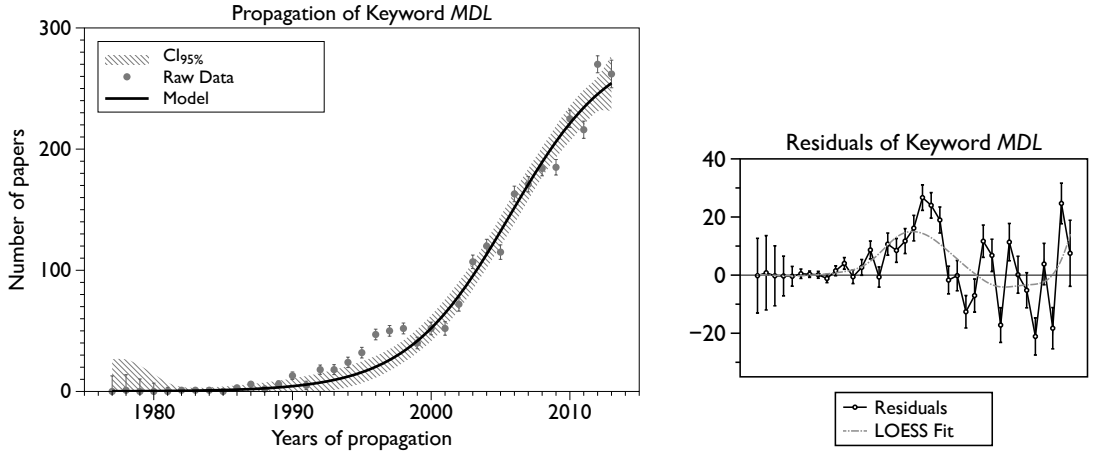
The resample produced $\bar{R}^2 = 0.947$ and $R_0 = 1.924$.

In case of the keyword "MANCOVA" and the *SEIR* model, the $\bar{R}^2 = 0.960$ and



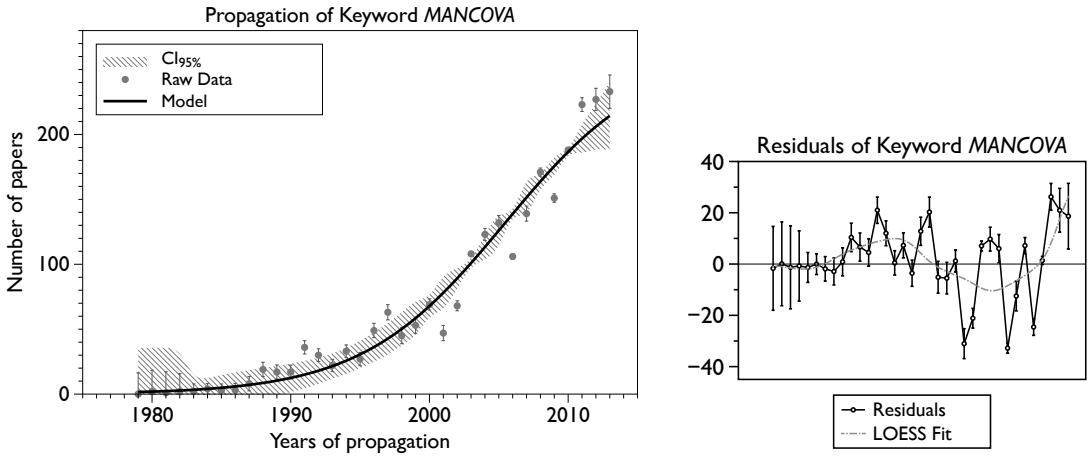
(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars *(b)* Residuals, Standard Error bars, and LOESS curve

Figure (6.44) Data and *SEIRE* model of keyword "Minimum Description Length."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars *(b)* Residuals, Standard Error bars, and LOESS curve

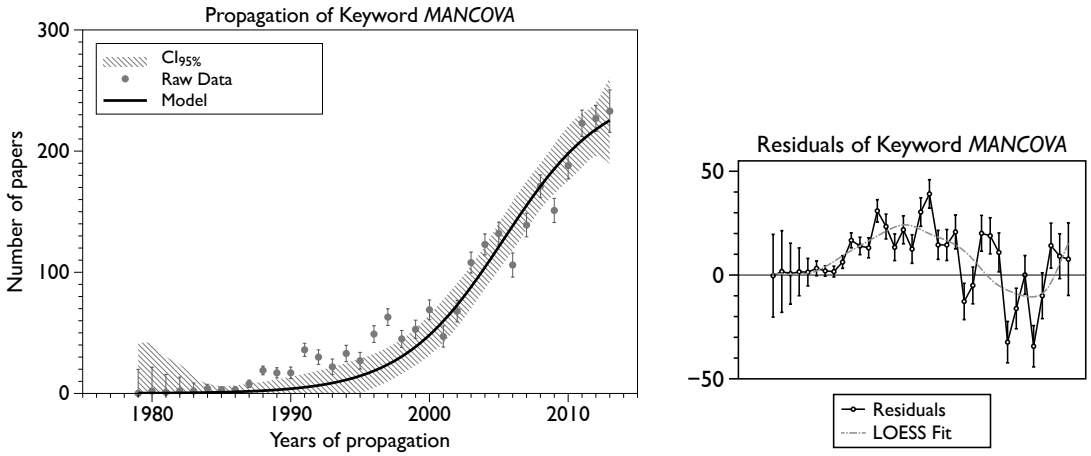
Figure (6.45) Data and *SEIR* model of keyword "Minimum Description Length."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

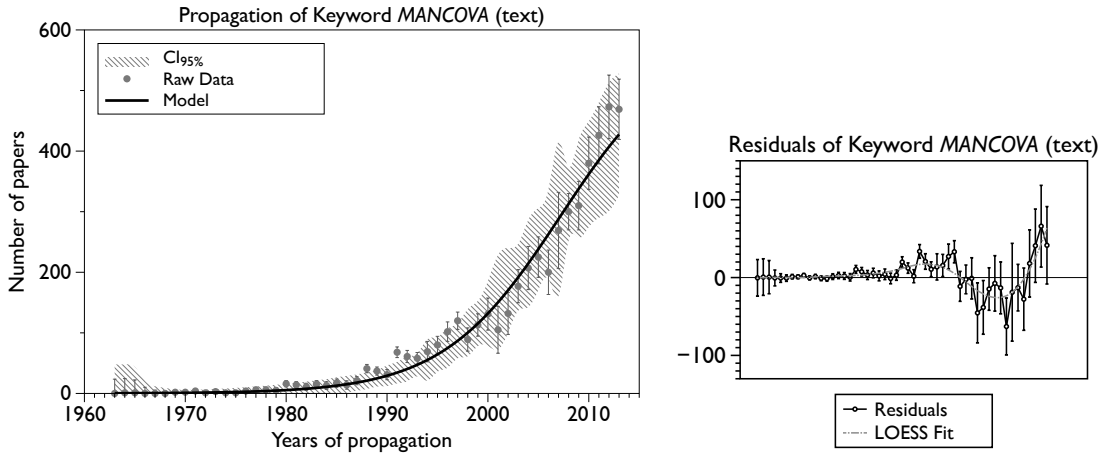
Figure (6.46) Data and *SEIRE* model of keyword "MANCOVA."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.47) Data and *SEIR* model of keyword "MANCOVA."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.48) Data and *SEIRE* model of keyword "Multivariate Analysis Of Covariance."

$R_0 = 2.45578$ (Figure 6.47).

The resample produced $\bar{R}^2 = 0.951$ and $R_0 = 1.51777$.

In case of the keyword "Multivariate Analysis Of Covariance" and the *SEIRE* model, the $\bar{R}^2 = 0.978$ and $R_0 = 2.663$ (Figure 6.48).

The resample produced $\bar{R}^2 = 0.983$ and $R_0 = 2.798$.

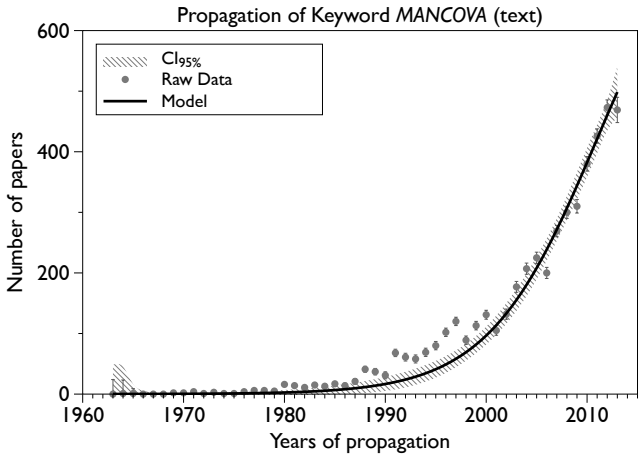
In case of the keyword "Multivariate Analysis Of Covariance" and the *SEIR* model, the $\bar{R}^2 = 0.978$ and $R_0 = 1.61895$ (Figure 6.49).

The resample produced $\bar{R}^2 = 0.970$ and $R_0 = 1.80018$.

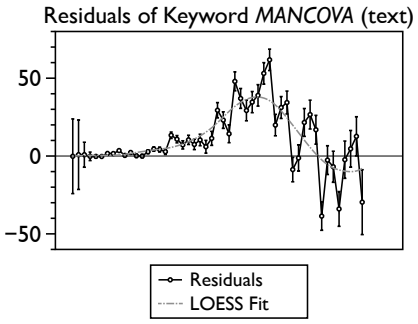
In case of the keyword "MySQL" and the *SEIRE* model, the $\bar{R}^2 = 0.968$ and $R_0 = 437.661$ (Figure 6.50).

The resample produced $\bar{R}^2 = 0.969$ and $R_0 = 62.348$.

In case of the keyword "MySQL" and the *SEIR* model, the $\bar{R}^2 = 0.974$ and $R_0 =$

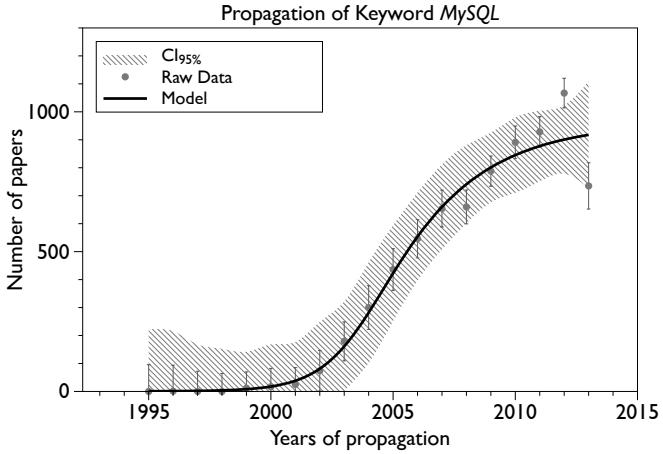


(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

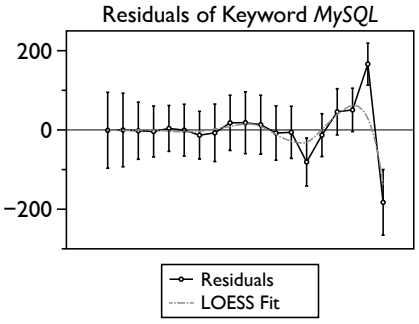


(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.49) Data and *SEIR* model of keyword "Multivariate Analysis Of Covariance."

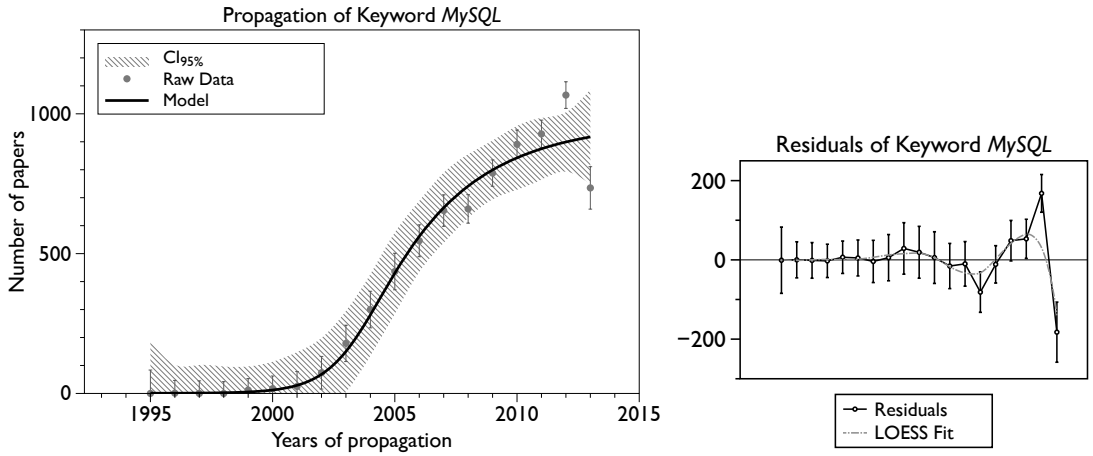


(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars



(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.50) Data and *SEIRE* model of keyword "MySQL."



(a) Data, $SEIR$ model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.51) Data and $SEIR$ model of keyword "MySQL."

28.977 (Figure 6.51).

The resample produced $\bar{R}^2 = 0.974$ and $R_0 = 416.574$.

In case of the keyword "Organic Light Emitting Diode" and the $SEIRE$ model, the $\bar{R}^2 = 0.991$ and $R_0 = 2527.540$ (Figure 6.52).

The resample produced $\bar{R}^2 = 0.984$ and $R_0 = 530.862$.

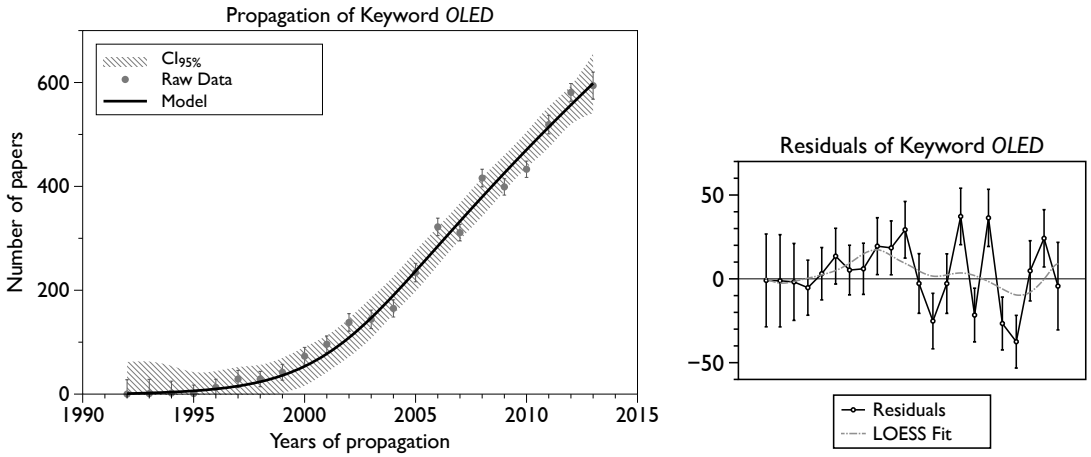
In case of the keyword "Organic Light Emitting Diode" and the $SEIR$ model, the $\bar{R}^2 = 0.980$ and $R_0 = 483.761$ (Figure 6.53).

The resample produced $\bar{R}^2 = 0.979$ and $R_0 = 690.111$.

In case of the keyword "Quantum Cryptography" and the $SEIRE$ model, the $\bar{R}^2 = 0.897$ and $R_0 = 1.547$ (Figure 6.54).

The resample produced $\bar{R}^2 = 0.909$ and $R_0 = 2.128$.

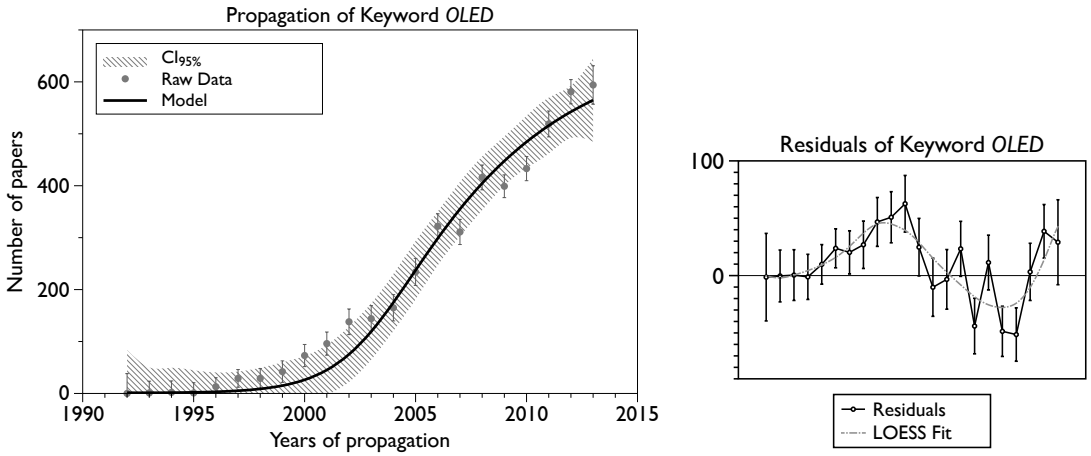
In case of the keyword "Quantum Cryptography" and the $SEIR$ model, the $\bar{R}^2 =$



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

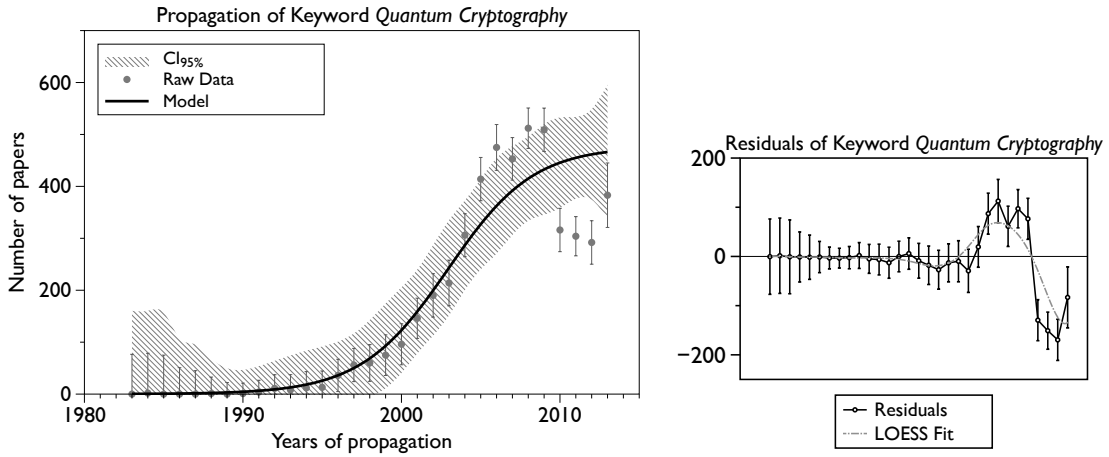
Figure (6.52) Data and *SEIRE* model of keyword "Organic Light Emitting Diode."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

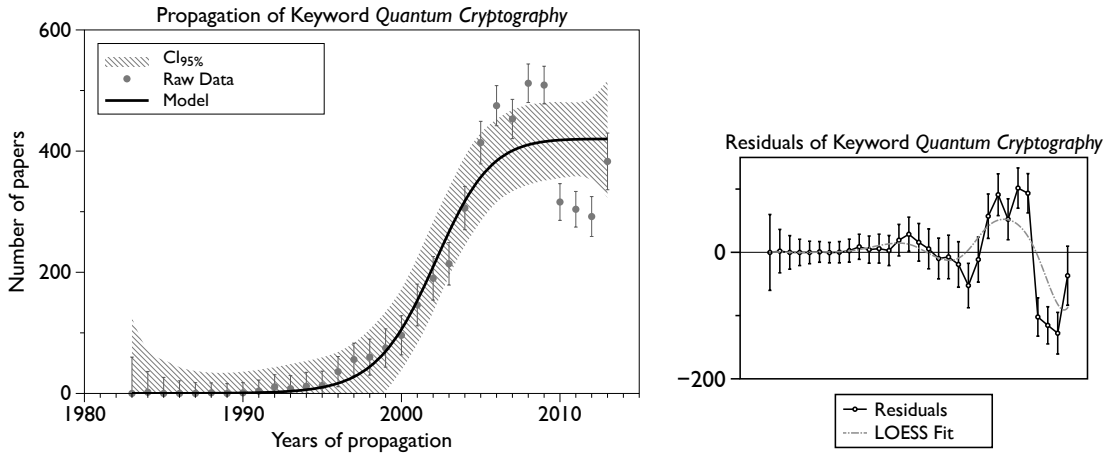
(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.53) Data and *SEIR* model of keyword "Organic Light Emitting Diode."



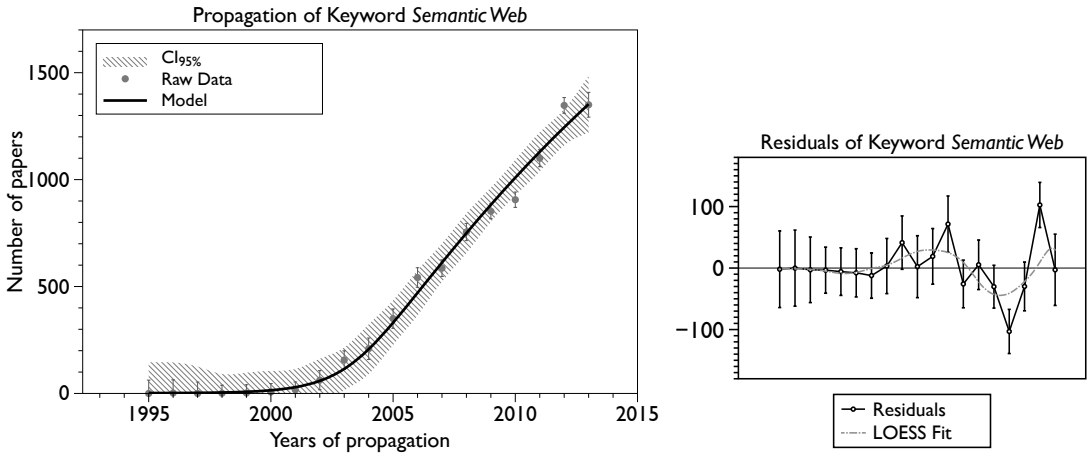
(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars *(b)* Residuals, Standard Error bars, and LOESS curve

Figure (6.54) Data and *SEIRE* model of keyword "Quantum Cryptography."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars *(b)* Residuals, Standard Error bars, and LOESS curve

Figure (6.55) Data and *SEIR* model of keyword "Quantum Cryptography."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.56) Data and *SEIRE* model of keyword "Semantic Web."

0.936 and $R_0 = 0.89989$ (Figure 6.55).

The resample produced $\bar{R}^2 = 0.946$ and $R_0 = 7.51578$.

In case of the keyword "Semantic Web" and the *SEIRE* model, the $\bar{R}^2 = 0.991$ and $R_0 = 1721.620$ (Figure 6.56).

The resample produced $\bar{R}^2 = 0.986$ and $R_0 = 934.185$.

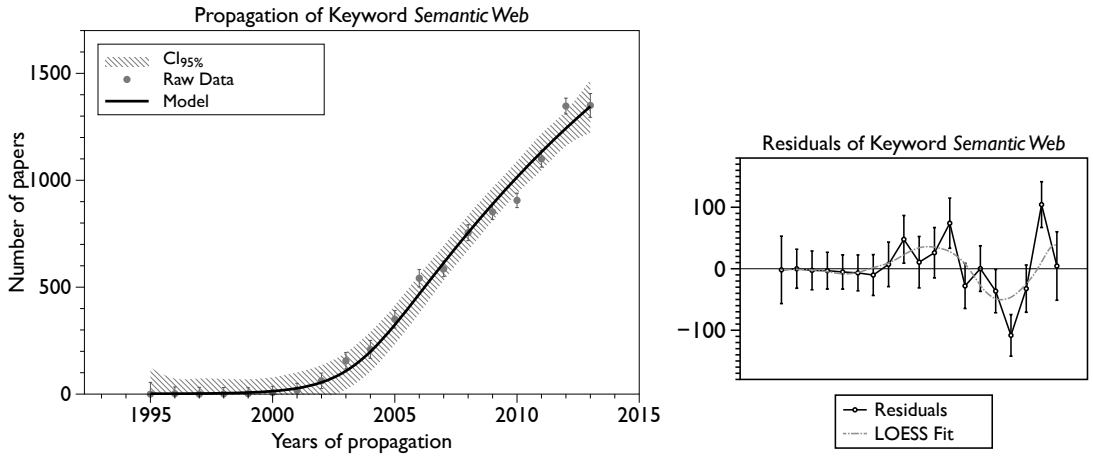
In case of the keyword "Semantic Web" and the *SEIR* model, the $\bar{R}^2 = 0.992$ and $R_0 = 1090.85$ (Figure 6.57).

The resample produced $\bar{R}^2 = 0.993$ and $R_0 = 1120.65$.

In case of the keyword "Smart Grid" and the *SEIRE* model, the $\bar{R}^2 = 0.995$ and $R_0 = 3861.950$ (Figure 6.58).

The resample produced $\bar{R}^2 = 0.995$ and $R_0 = 3841.440$.

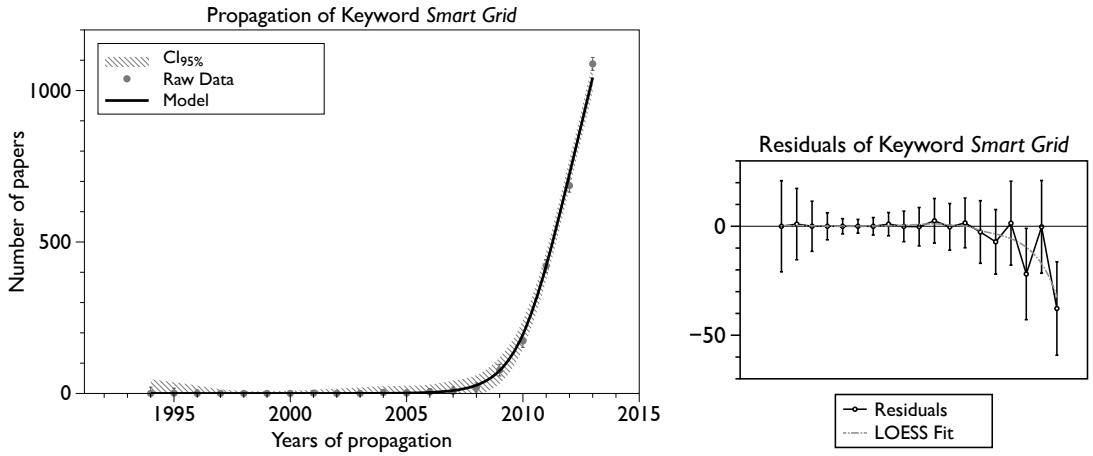
In case of the keyword "Smart Grid" and the *SEIR* model, the $\bar{R}^2 = 0.998$ and



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

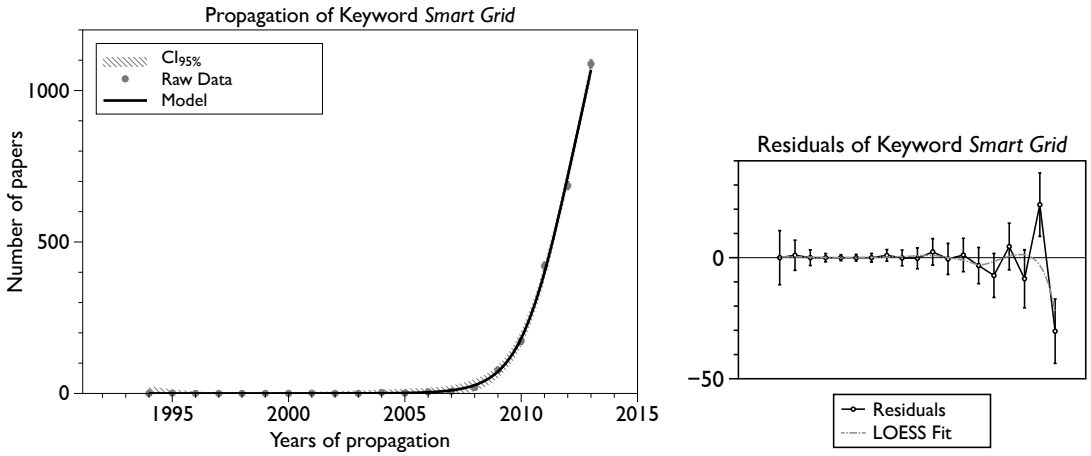
Figure (6.57) Data and *SEIR* model of keyword "Semantic Web."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.58) Data and *SEIRE* model of keyword "Smart Grid."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.59) Data and *SEIR* model of keyword "Smart Grid."

$R_0 = 1860$ (Figure 6.59).

The resample produced $\bar{R}^2 = 0.998$ and $R_0 = 2072.66$.

In case of the keyword "Spintronics" and the *SEIRE* model, the $\bar{R}^2 = 0.973$ and $R_0 = 303.571$ (Figure 6.60).

The resample produced $\bar{R}^2 = 0.975$ and $R_0 = 138.865$.

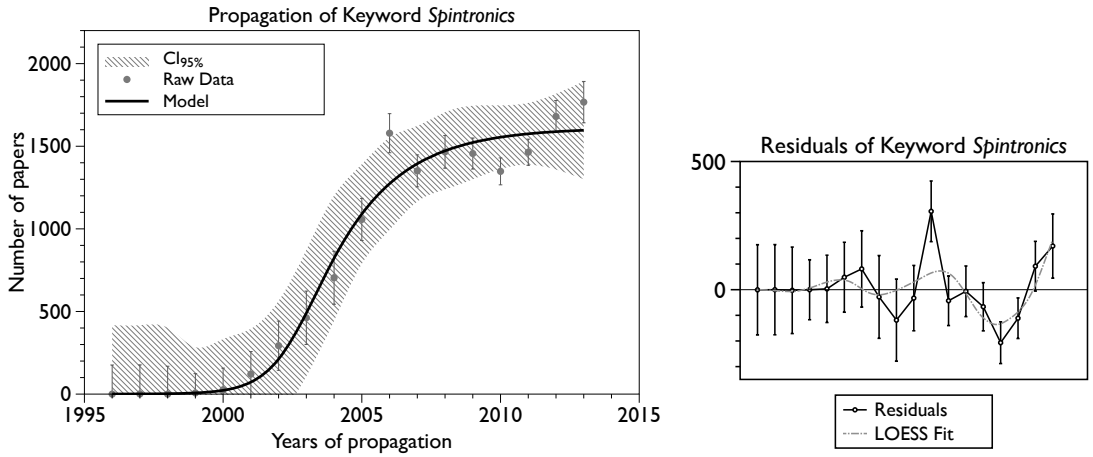
In case of the keyword "Spintronics" and the *SEIR* model, the $\bar{R}^2 = 0.979$ and $R_0 = 205.429$ (Figure 6.61).

The resample produced $\bar{R}^2 = 0.981$ and $R_0 = 136.036$.

In case of the keyword "Stem Cell Research" and the *SEIRE* model, the $\bar{R}^2 = 0.991$ and $R_0 = 4.750$ (Figure 6.62).

The resample produced $\bar{R}^2 = 0.969$ and $R_0 = 4.427$.

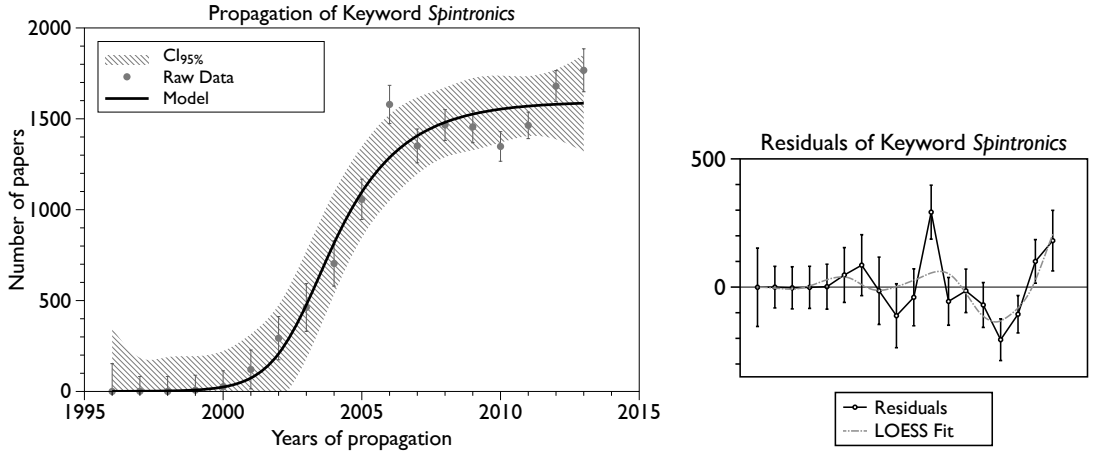
In case of the keyword "Stem Cell Research" and the *SEIR* model, the $\bar{R}^2 = 0.993$



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

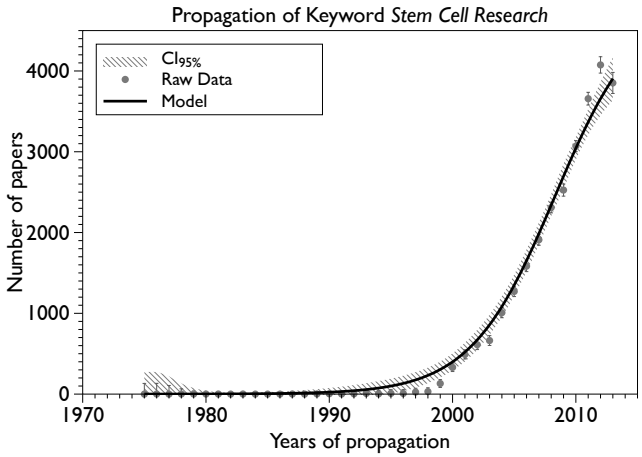
Figure (6.60) Data and *SEIRE* model of keyword "Spintronics."



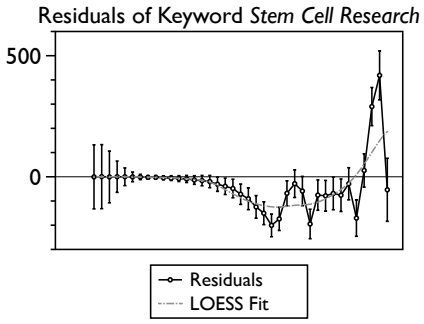
(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.61) Data and *SEIR* model of keyword "Spintronics."

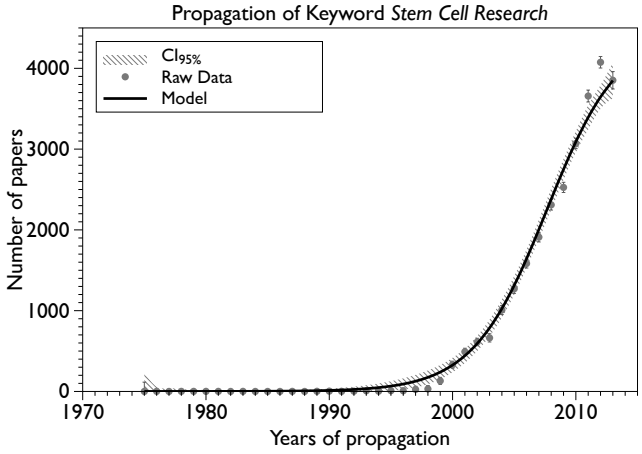


(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

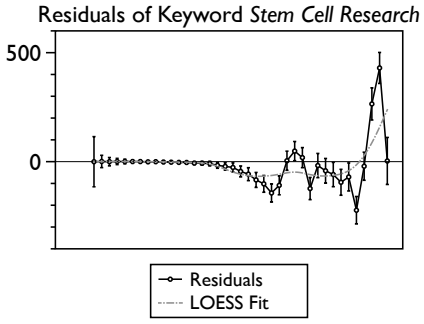


(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.62) Data and *SEIRE* model of keyword "Stem Cell Research."

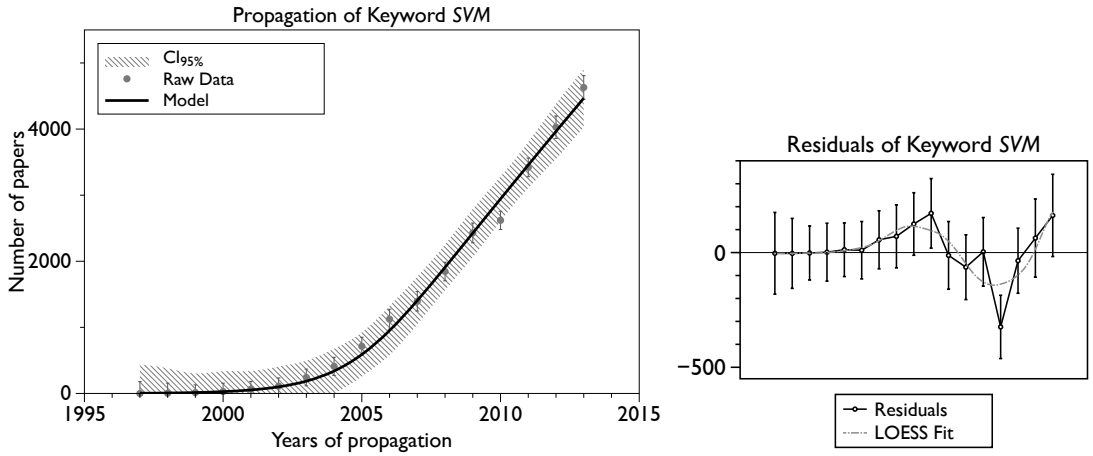


(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars



(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.63) Data and *SEIR* model of keyword "Stem Cell Research."



(a) Data, $SEIRE$ model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.64) Data and $SEIRE$ model of keyword "Support Vector Machine."

and $R_0 = 1.28947$ (Figure 6.63).

The resample produced $\bar{R}^2 = 0.982$ and $R_0 = 1.53905$.

In case of the keyword "Support Vector Machine" and the $SEIRE$ model, the $\bar{R}^2 = 0.992$ and $R_0 = 6379.470$ (Figure 6.64).

The resample produced $\bar{R}^2 = 0.992$ and $R_0 = 6285.450$.

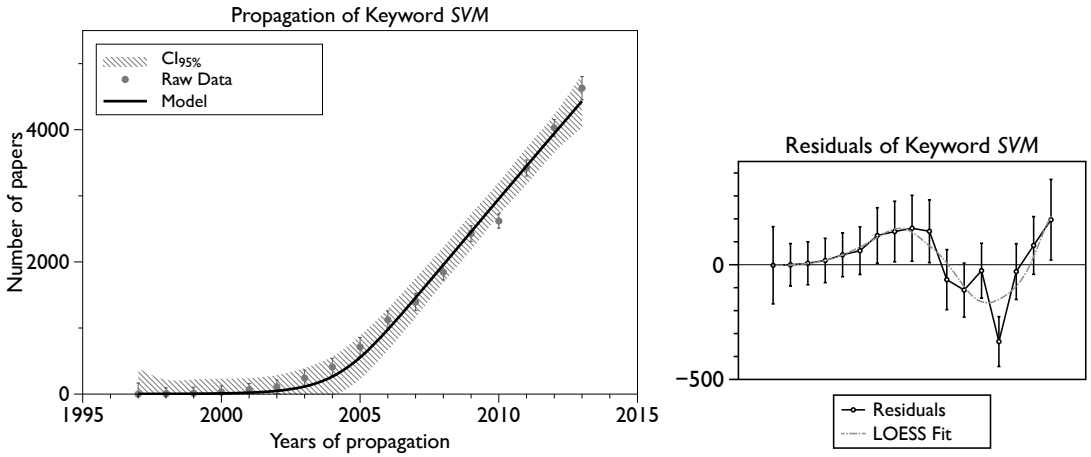
In case of the keyword "Support Vector Machine" and the $SEIR$ model, the $\bar{R}^2 = 0.988$ and $R_0 = 3714.98$ (Figure 6.65).

The resample produced $\bar{R}^2 = 0.979$ and $R_0 = 2131.93$.

In case of the keyword "System on a Chip" and the $SEIRE$ model, the $\bar{R}^2 = 0.975$ and $R_0 = 1.425$ (Figure 6.66).

The resample produced $\bar{R}^2 = 0.969$ and $R_0 = 6.085$.

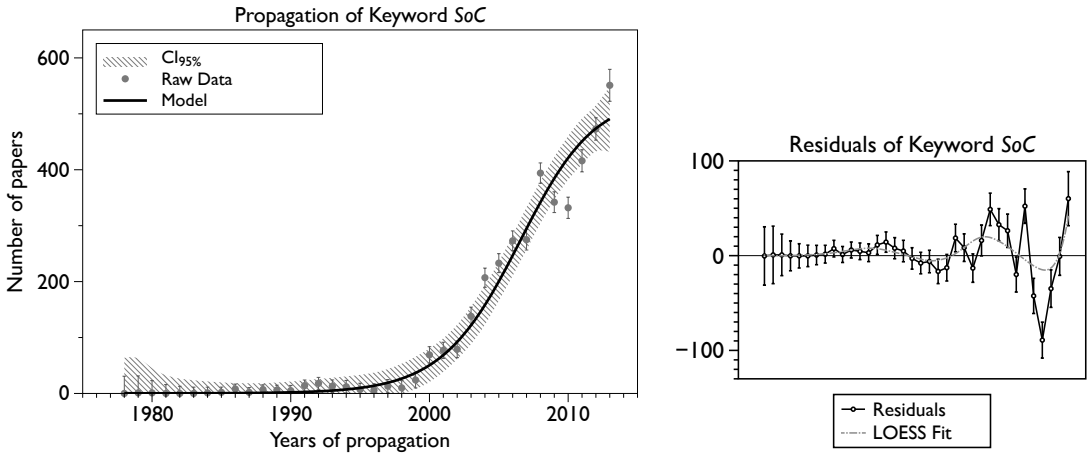
In case of the keyword "System on a Chip" and the $SEIR$ model, the $\bar{R}^2 = 0.980$ and



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

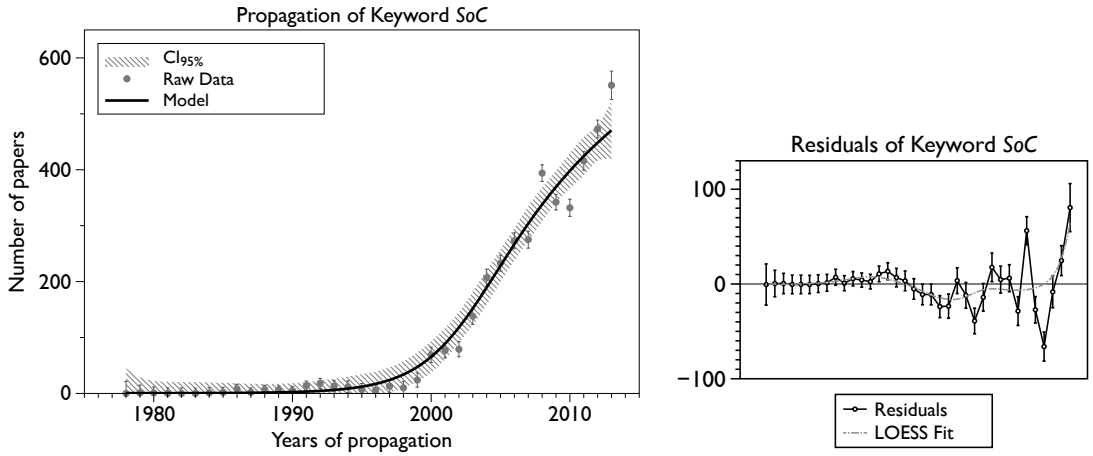
Figure (6.65) Data and *SEIR* model of keyword "Support Vector Machine."



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.66) Data and *SEIRE* model of keyword "System on a Chip."



(a) Data, $SEIR$ model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.67) Data and $SEIR$ model of keyword "System on a Chip."

$R_0 = 29.7624$ (Figure 6.67).

The resample produced $\bar{R}^2 = 0.983$ and $R_0 = 1913.07$.

In case of the keyword "Systems Integration" and the $SEIRE$ model, the $\bar{R}^2 = 0.953$ and $R_0 = 4.096$ (Figure 6.68).

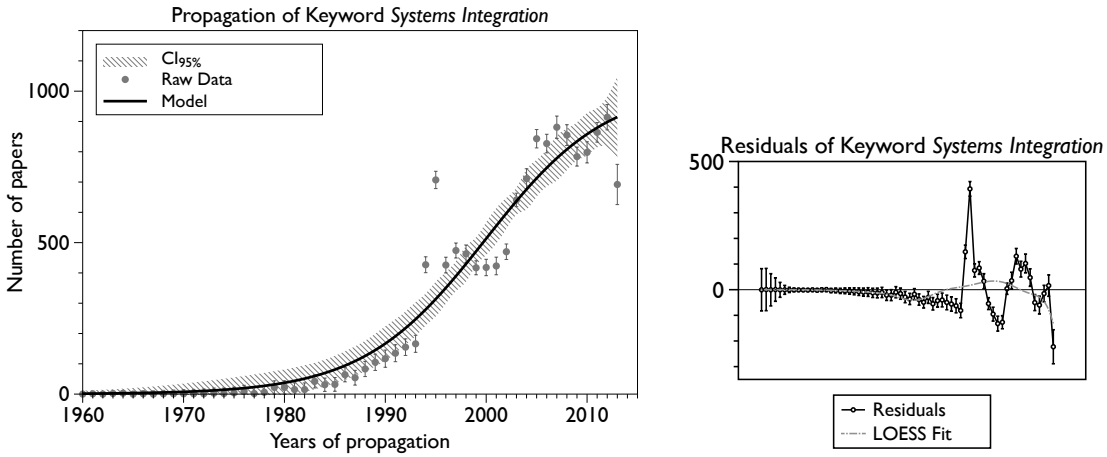
The resample produced $\bar{R}^2 = 0.894$ and $R_0 = 5.235$.

In case of the keyword "Systems Integration" and the $SEIR$ model, the $\bar{R}^2 = 0.939$ and $R_0 = 8126.76$ (Figure 6.69).

The resample produced $\bar{R}^2 = 0.933$ and $R_0 = 7931.4$.

Both models show high tracking abilities for the keywords presented here. Other propagation characteristics that do not resemble epidemiological attributes do also exist. Both models fail in that case.

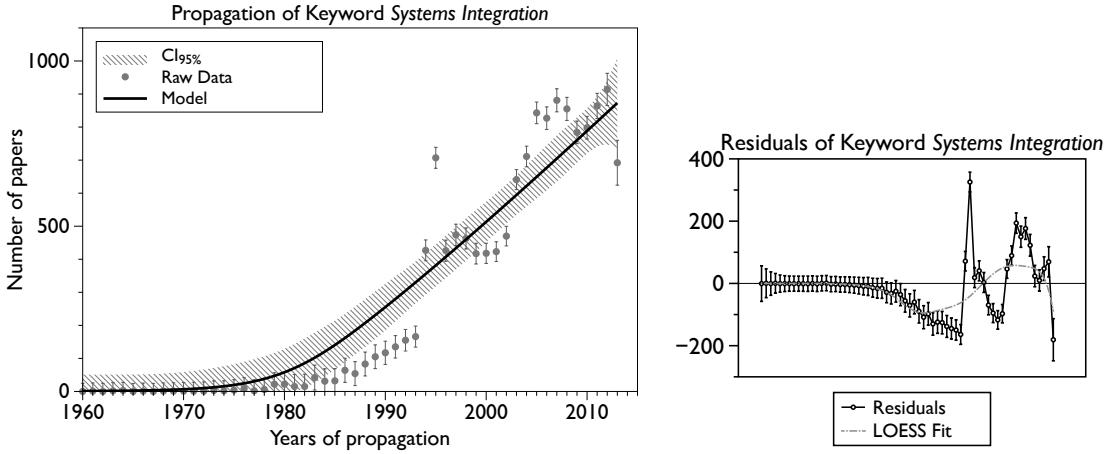
In some research the Basic Reproductive Rate can give insight into characteristics that



(a) Data, *SEIRE* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.68) Data and *SEIRE* model of keyword "Systems Integration."



(a) Data, *SEIR* model, 95% Confidence Interval, and Standard Error bars

(b) Residuals, Standard Error bars, and LOESS curve

Figure (6.69) Data and *SEIR* model of keyword "Systems Integration."

follow an outbreak with $R_0 > 1$ or demise with $R_0 < 1$ of a contagion agent. This was not the case for knowledge diffusion. One reason for this is judged to be the overall propagation time. Knowledge that formed epidemiological behavior after three or more decades with a relatively small number of published papers (hundreds to low thousands) had R_0 values smaller than 1 (e.g. Figure 6.31 or 6.43).

In cases like *Systems Integration* (Figure 6.68 and 6.69), the implications of R_0 are quite different. The $SEIR$ model does not track the propagation properly and the values suggest a very high propagation potential. As in classical epidemiology, many of these values are open to interpretation. However, general inference on the propagation characteristics can be made. Beyond that, the $SEIRE$ model provides a ratio of direct knowledge transfer and knowledge transfer from hidden knowledge-holders. This ration (β:δ) is reflected by the $SEIRE$ model's Basic Reproduction Rate R_0 (Equation 6.11).

CHAPTER 7

CONCLUSIONS

The tracking of knowledge in communities is an intriguing field with the potential to understand knowledge transfer and diffusion. Being able to track the propagation of knowledge in scientific publication is one step to understanding diffusion characteristics.

From all of the examined data sets can be seen that the propagation of knowledge in scientific publications takes several decades to reach epidemiological behavior. Recent discoveries show faster initial adoption, but display similar growth in publication numbers thereafter.

The author has been able to successfully track the spread of knowledge over several decades using the original, unaltered $SEIR$ model. Further improvements have been implemented by modifying the compartmental model to better fit the knowledge propagation in scientific publications with the $SEIRE$ model. 17 keywords from fields of data

mining, statistical analysis, Soft Computing, and engineering could be tracked with high accuracy. The list of keywords was extended to 34 (out of a total of 88) to be able to conduct an in-depth comparison. A more robust approach and improved tracking capabilities demonstrate that causality approach in combination with the foundations in place can lead to further inquiries. Although successful, some knowledge propagation characteristic problems remain. One unanswered question is that of changing topics over time, i.e. evolution of knowledge and could be addressed with the Price Equation as used in the paper from Henrich (2002).

The parameter-wise clustering of the keywords gave an insight to the long-term propagation development. Taking into consideration the size of the current data set, a classification into two groups proved to be the most informative. With larger data sets, however, grouping into more classes could demonstrate further unique shared attributes. The author assumes that this knowledge discovery may be used to assist in detecting and assessing evolutionary characteristics.

One of the conclusions is that there is more potential to extend the deterministic models. knowledge propagation specific features and reasoning could improve tracking the existing diffusion process as well as fit the ones that could not be tracked with the simple $SEIR$ or the extended $SEIRE$ model. The potential is limited however and other methodologies such as agent modeling or the Price Equation might be more apt in addressing specific problems.

Some of these limitations of the deterministic models have been addressed. The nature of ordinary differential equations might prohibit implementing the transfer of knowledge in an evolutionary sense, i.e. change over time. Other methods as described in the

introductory part of this report, especially agent simulation seems fit to be used as dynamic elements. Incorporating the knowledge gained from this research, cellular automata and diffusion equations could be considered to address unanswered dynamics.

Although the generic $SEIR$ model demonstrated good tracking ability, the comparison of the models shows the necessity of extending the basic epidemiological model using causality as a guiding fundament. This is also well supported by previous research. The limits with deterministic models are visible by minor gain in causality and a slight increase in performance but with the downside of increased complexity. The clear causality of the $SEIRE$ model is favored over the minimal gain of performance and an ambiguous meaning when incorporating the K (hidden knowledge-holders with influence on the propagation of knowledge, but not directly through renewed publications) compartment in the $SEIRK$ model.

The large data sets of each database in Japan, China, and worldwide academic publications allowed for high performance in fitting the epidemiological model to the data. Each of the cultural categories could be classified using Principal Component Analysis on both combined parameters and single parameters in combination with k-means clustering. The use of high-volume databases of scientific publications opened the possibility for more detailed examinations of keywords from science and engineering.

With the use of the basic deterministic epidemiological model ($SEIR$), this research indicates that the parameters of this model can give quantitative insight into existing qualitative cultural models. In this paper the authors showed that the collective parameters of the $SEIR$ model can be used to distinguish cultural propagation characteristics in Japan, China, and worldwide for five methodologies in Soft Computing. The identification of

these categories using data analysis gives basis for further investigation. Hofstede's dimensions, Hall's cultural factors, or Schwartz' human values could be used to better understand the epidemiological propagation dynamics of scientific knowledge. Parameters of the deterministic model that are connected to a cultural dimension could also help to improve the predictability of the epidemiological model.

For sufficient data points, both the epidemiological $SEIR$ model as well as the novel $SEIRE$ model track knowledge with high accuracy. The causality implications show that the influence of hidden knowledge-holders is quite high and can be used to explain some of the variance unanswered by the generic model.

The overall conclusion of tracking knowledge propagation in scientific publications using deterministic epidemiological models is that the $SEIRE$ model is the most suitable method (compared to the generic $SEIR$ model, the $SEIRK$, and $SEIREK$ model). The causality of the model, emphasizing individuals who possess knowledge and exercise influence on fellow researchers to spread it, supports the performance increase in the IEEE Xplore Digital Library data set as well as superior tracking in the Scirus data base.

GLOSSARY

AdaBoost refers to a Machine Learning algorithm and stands for Adaptive Boosting. vi, 50, 62, 84, 85

AIC short for Akaike Information Criterion, is a method to test the goodness of fit of a statistical model to the data. 50, 62, 129

ANCOVA Analysis of Covariance is a combination of Analysis of Variance (ANOVA) and regression analysis because the model contains both quantitative and qualitative independent variables. The idea is to enhance the ANOVA model by adding one or more quantitative independent variables that are related to the dependent variable. vi, vii, 50, 84–86, 133

ANOVA Analysis of Variance is a statistical technique that isolates and assesses the contribution of categorical independent variables to the variance of the mean of a continuous dependent variable. The observations are classified according to their

categories for each of the independent variables, and the differences between the categories in their mean values on the dependent variable are estimated and tested for statistical significance. 6, 43, 50, 62, 132, 133

AR Augmented Reality refers to the merging of synthetic sensory information into a user's perception of a real environment (Rohaya et al., 2012). vii, 50, 88, 89

ASIC is an Application Specific Integrated Circuit and similar to SoC but often focused on a specific task, usually with a more complex application than FPGA. vii, 50, 64, 86, 87

Backpropagation was used as a keyword as it is a common technique to train Neural Networks. vii, 50, 62, 63, 66, 89, 90

Basic Reproductive Rate (R_0) Often called *Basic Reproduction Number* in scientific papers, this is a measure of the number of infections produced, on average, by an infected individual in the early stages of an epidemic, when virtually all contacts are susceptible. 4, 22–24, 30, 35, 59, 76, 83, 84, 119, 121

Bayesian Network is a probabilistic graphical model, often used in Soft Computing. *Belief Network* was as well used as a representative keyword. 39, 50, 51, 62, 66, 74, 81, 82, 133, 136

BCI Brain Computer Interface is the communication gateway between the brain and a computational device. vii, 50, 92, 93

BIC Bayesian Information Criterion is, similar to the *Akaike Information Criterion*, a measure of goodness of fit of a statistical model. vii, 50, 63, 90, 91

C4.5 "The decision tree program C4.5 and its successor C5.0 were devised by Ross Quinlan over a 20-year period beginning in the late 1970s"—Witten and Frank (2011). 50, 63

CCA Canonical Correlation Analysis is used in statistics to assess cross-covariance matrices. vii, 50, 64, 92–94

χ^2-Test was considered due to the importance in statistical hypothesis testing. "The presence of an association between two categorical variables, such as exposure and outcome, can also be tested using the chi-square (χ^2) test statistic. The χ^2 test is used to determine if the observed data in a 2×2 table are statistically significantly different from what would be expected, given the row and column totals for the table. Chi-square testing can be performed for a single 2×2 table, or it can be used to test for overall association when performing stratified analysis, which is discussed in detail below."—Boslaugh (2008). 50, 63

contagion agent Biological and non-biological infectious agent. "Epidemics are most often caused by infectious agents such as viruses or bacteria, but other causes are possible, including chemical exposure and physical conditions such as extreme heat or cold"—Boslaugh (2008). 1, 7, 11, 19, 29, 74, 78, 121

CPLD Complex Programmable Logic Device features a wider application range than FPGA in programming logical devices. vii, 50, 94, 95

Decision Tree stands for a category of methods of decision support tools. 50, 62, 63, 66

Document Clustering was chosen as it is a text clustering concept. vii, 50, 63, 94, 96

endemic See epidemic. 1, 17, 21, 59, 65

epidemic "An epidemic is a marked increase in the number of cases of a disease relative to the expected number of cases. Epidemic disease is sometimes contrasted with endemic disease, which is the expected or usual incidence of disease in a location. While the term endemic is typically confined to infectious diseases, the term epidemic is more widely used. Endemic can refer to either the usually observed rate of disease or simply the fact that a disease is present in a locale. For example, hantavirus is endemic to many parts of the United States. A rate of disease that is endemic on one country would constitute an epidemic if it occurred in a country where the disease is ordinarily less common."—Boslaugh (2008). 1, 13, 17, 20–22, 28, 34, 43, 59, 76, 130

epidemiology The study of the occurrence and distribution of health-related states or events in specified populations, including the study of the determinants influencing such states, and the application of this knowledge to control the health problems. 1, 2, 12, 13, 15, 17–20, 27, 28, 55, 64, 121

Evolutionary Algorithm as a subcategory of evolutionary computation is used mainly in artificial intelligence. There are a manifold of algorithms falling under this category.

From literature, the most used methodologies were selected as follows: *Evolutionary Algorithm*, *Gene Expression Programming*, *Genetic Algorithm*, *Genetic Programming*, *Evolutionary Programming*, *Evolution Strategy*, *Memetic Algorithm*, *Differential Evolution*, *Neuroevolution*, and *Learning Classifier System*. 39, 50, 51, 62, 66, 74, 81, 82, 131, 136

FPGA means *Field-Programmable Gate Array* and is a uniquely designed integrated circuit to perform an exact task with optimum performance. 50, 62, 64, 66, 128, 129

Fractal refers to mathematical sets that are self-similar but non-trivial. vii, 50, 97, 98

Fuzzy Logic evolved from the fuzzy set theory of Lofti Zadeh and is a form of probabilistic reasoning that allows using approximate values instead of exact ones for computation. Fuzzy logic has been used in various fields and applied in the industry as well. vii, 6, 39, 50, 51, 53, 61, 62, 65, 69, 74, 81, 82, 97–99, 136

Genetic Algorithm is one of the most used methods of the subclass of *Evolutionary Algorithms*. vii, viii, 39, 50, 51, 61, 62, 66, 69, 74, 81, 82, 99, 100, 131, 136

GPGPU General Purpose Computing On Graphics Processing Units is sometimes referred to as *GPGP* or GP^2U. 50

HDL denotes *Hardware Description Language* and is a category of programming in digital logic. 50, 62, 64, 66

Hebbian Learning is a learning process for training Artificial Neural Networks in information science. 50, 63

HMM Hidden Markov Model was considered as computational sequence analysis and generating probabilistic models. 50, 62, 63, 65

HTML is the most popular online programming language and both the abbreviation *HTML* and *Hypertext Markup Language* were analyzed for this paper. 50, 51, 62, 63

HTML5 is the newest version of HTML to date. 50

k-means is the most frequently-used nonhierarchical clustering technique, which is inspired by the principles of analysis of variance. In fact, it may be thought of as an analysis of variance "in reverse." If the number of clusters is fixed as k, the algorithm will start with k random clusters and then move objects between them with the goals of minimizing variability within clusters and maximizing variability between clusters. viii, 3, 39, 50, 63, 65, 66, 99, 101, 125

Kansei Engineering is sometimes referred to as *Affective Engineering* or *Emotional Engineering*. viii, 50, 102, 103

knowledge Knowledge as considered in this thesis represents a research method, methodology, process, algorithm, or the like that scientists use in their research and development. v, vi, xi, 1–14, 19, 27–29, 35, 39–41, 43, 44, 48, 49, 55, 56, 58, 59, 62, 64, 65, 71–75, 77, 80, 81, 83, 121, 123–126

LDA Linear Discriminant Analysis is related to ANOVA and Principal Component Analysis and is used for pattern recognition, Machine Learning, and statistical analysis. 50, 63

LED Light Emitting Diode is a light source made from semiconductors. viii, 50, 102–104

LSI is the predecessor of VLSI and stands for *Large Scale Integration*. 50, 64

Machine Learning is a term including categories of algorithms such as SVM, genetic programming, Bayesian Networks, Neural Networks, and other, and was picked for analysis. 50, 62, 63, 66, 127, 132, 136, 169, 171

MANCOVA Multivariate Analysis Of Covariance. See MANOVA and ANCOVA. viii, 50, 104, 106–108

MANOVA Multivariate Analysis Of Variance is a statistical technique used extensively in all types of research. It is the same thing as an Analysis of Variance (ANOVA), except that there is more than one dependent or response variable. The mathematical methods and assumptions of MANOVA are simply expansions of ANOVA from the univariate case to the multivariate case. 50, 133

MDL is short for *Minimum Description Length*, used for inductive inference in information theory. viii, 50, 63, 104, 105

MDP Markov Decision Process is a framework for decision-making modeling. 50, 63

MDS Multidimensional Scaling is often used for visual representation of information and is a subfield of ordination in multivariate analysis. 50, 64

MySQL is one of the most widely-used open-source electronic databases. viii, 50, 107–109, 136

Neural Network refers to Artificial Neural Networks. vi, 39, 50, 51, 61, 62, 66, 69, 74, 77, 81, 82, 128, 131, 133, 136, 160, 170, 171, 173

OLED Organic Light Emitting Diode is an LED with an organic film as the light emitter. viii, 50, 109, 110

pandemic "An epidemic occurring worldwide, or over a very wide area, crossing international boundaries, and usually affecting a large number of people"—Last and Abramson (2001). 1, 21

PCA The application range of Principal Component Analysis covers dimension reduction of data sets, pattern recognition, and exploratory data analysis. 4, 6, 39, 50, 61, 63, 65, 66, 69, 125, 132

Perceptron was used to refer to a supervised classification algorithm. 50, 62, 63, 66

Quantum Cryptography is a new field of cryptography taking the quantum mechanical effect of quantum computation and communication into account. viii, 50, 109, 111

SEIR Basic deterministic epidemiological model with four compartments—susceptible, exposed, infected, recovered. v–ix, xi, 2–4, 11, 13, 15, 19, 24, 28–30, 34, 55, 58–60, 64, 74–79, 81, 83–121, 123–126

SEIRE Revised deterministic model with four compartments—susceptible, exposed (learning and paper writing), infected (paper publication), recovered (change research

field, etc.). Once recovered, researchers can enter the E compartment again, indicating another publication with the same knowledge. v–ix, xi, 3, 4, 12, 14, 15, 75–121, 123–126

$SEIREK$ Revised deterministic model with five compartments—susceptible, exposed (learning and paper writing), infected (paper publication), recovered (change research field, etc.), hidden knowledge-holders. This model combines the $SEIRE$ and $SEIRK$ model. xi, 3, 79, 81, 82, 126

$SEIRK$ Revised deterministic model with five compartments—susceptible, exposed (learning and paper writing), infected (paper publication), recovered (change research field, etc.), hidden knowledge-holders. Once researchers enter the K compartment, they stop publishing papers with the same knowledge but still influence on the propagation process. vi, xi, 3, 79, 81, 82, 125, 126

Semantic Web A key approach is to use data mining techniques that mediate between the different levels of meta data and to use semantic web technology to store and transport the knowledge (May and Saitta, 2010). viii, 50, 112, 113

SIR Basic deterministic epidemiological model with three compartments—susceptible, infected, recovered. v, 2, 11, 13, 20–24, 31, 32

Smart Grid is an electrical grid that uses ICT (Information and Communications Technology) to update information for optimizing load and usage. viii, 50, 112–114

SoC is short for *System on a Chip* and is a concept of merging different integrated circuits to a system on a chip to form one token to perform as a cluster of components. ix,

50, 62, 64, 66, 117–119, 128

Soft Computing Soft Computing in computer science refers to algorithms addressing uncertainty and inexact solutions. In this research, five most widely applied methodologies were used, namely Fuzzy Logic, Neural Network, Evolutionary Algorithm, Bayesian Network, and Genetic Algorithm. xi, 3, 4, 39, 49, 51, 53, 74, 82, 124, 125, 128

SOM is short for *Self Organizing Map* and is a subclass of artificial Neural Networks. 50, 63

Spintronics is an emerging technology and studies the role of electron spins. ix, 50, 114, 115

SQL was chosen as one of the standards in database programming language.

In later analysis, *SQL* was discarded and *MySQL* included. As SQL is one of many programming languages, it was not considered as a leading technology, whereas MySQL is one of the most used Open Source database formats to date. 63

SRAM or *Static Random-Access Memory*, is a type of semiconductor based memory block. 64

Stem Cell Research and related topics include *Stem Cell Treatment* and *Stem Cell Therapy*. ix, 50, 114, 116

SVM Support Vector Machine is a subfield of Machine Learning—a supervised learning model—"Support Vector Machines are based on the Statistical Learning Theory

concept of decision planes that define decision boundaries. A decision plane ideally separates objects having different class memberships."—Nisbet et al. (2009). ix, 50, 61, 63, 65, 117, 118, 133

ULSI stands for *Ultra Large Scale Integration*. 50, 173

VLSI stands for *Very Large Scale Integration*, the process of fusing a vast number of transistors on an integrated circuit. 6, 50, 61, 64, 66, 133, 161, 173

XML is short for *Extensible Markup Language*, is a commonly used file type for reading and writing textual data. XML was selected as the only file type in this research and is therefore an exception. Other file types such as CSV, DAT, and other had too few data points to consider for propagation tracking. 49, 50, 62

BIBLIOGRAPHY

Abdullah, S. and Wu, X. (2011). An epidemic model for news spreading on twitter. *IEEE International Conference on Tools with Artificial Intelligence*, pages 163–169.

Allen, B. (1982). A stochastic interactive model for the diffusion of information. *The Journal of Mathematical Sociology*, 8(2):265–281.

Alpaydin, E. (2004). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Baggio, R. (2006). The web graph of a tourism system. *Physica A: Statistical Mechanics and its Applications*, 379(2):727–734.

Bartholomew, D. J. (1973). *Stochastic models for social processes*. J. Wiley, New York.

Belen, S. (2008). The behaviour of stochastic rumours.

Bettencourt, L. M. A., Cintronarias, A., Kaiser, D., and Castillo-Chávez, C. (2006). The

power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, 364:513–536.

Bonissone, P. P. (1997). Soft computing: the convergence of emerging reasoning technologies. *Soft computing*, 1(1):6–18.

Boslaugh, S. (2008). *Encyclopedia of Epidemiology*. Sage Publications.

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (1978). *Statistics for experimenters*. Wiley-Interscience, 2nd edition.

Caselli, F. and II, W. J. C. (2001). Cross-country technology diffusion: The case of computers. *American Economic Review, American Economic Association*, 91(2):328–335.

Cavalli-Sforza, L. L. (1981). *Cultural transmission and evolution: a quantitative approach*. Princeton University Press.

Chowell, G., Fenimore, P., Castillo-Garsow, M., and Castillo-Chavez, C. (2003). Sars outbreaks in ontario, hong kong and singapore: the role of diagnosis and isolation as a control mechanism. *Journal of Theoretical Biology*, 224(1):1–8.

Coggon, D., Rose, G., and Barker, D. (1997). *Epidemiology for the Uninitiated*. BMJ Publications, London, 4 edition.

Cointet, J.-P. and Roth, C. (2007). How realistic should knowledge diffusion models be? *Journal of Artificial Societies and Social Simulation*, 10(3).

Cowan, R. and Jonard, N. (2001). Knowledge creation, knowledge diffusion and network structure. *Economics with Heterogeneous Interacting Agents*, 503(2001):327–343.

BIBLIOGRAPHY

da Fontoura Costa, L. and Baggio, R. (2009). The web of connections between tourism companies: Structure and dynamics. *Physica A: Statistical Mechanics and its Applications*, 388(19):4286–4296.

Daley, D. J. and Kendall, D. G. (1964). Epidemics and rumours. *Nature*, 204(4963):1118–1118.

Dickinson, R. E. and Pearce, C. E. M. (2003). Rumours, epidemics, and processes of mass action: Synthesis and analysis. *Mathematical and Computer Modelling*, 38(11-12):1157–1167.

Dietz, K. (1967). Epidemics and rumours: A survey. *Journal Of The Royal Statistical Society Series A General*, 130(4):505–528.

Dodds, P. S. and Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604.

Erumban, A. A. and de Jong, S. B. (2006). Cross-country differences in ICT adoption: A consequence of Culture? *Journal of World Business*, 41(4):302–314.

Fagerberg, J. and Verspagen, B. (2002). Technology-gaps, innovation-diffusion and transformation: an evolutionary interpretation. *Research Policy*, 31(8):1291–1304.

Ferguson, R. (2008). Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing. *Journal of Consumer Marketing*, 25(3):179–182.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52:399–433.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Frank, S. A. (1997). The price equation, fisher's fundamental theorem, kin selection, and causal analysis. *Evolution*, 51(6):1712.

Funkhouser, G. R. and McCombs, M. E. (1972). Predicting the diffusion of information to mass audiences. *Journal of Mathematical Sociology*, 2:121–130.

Gales, L. (2008). The role of culture in technology management research: National Character and Cultural Distance frameworks. *Journal of Engineering and Technology Management*, 25(1-2):3–22.

Gallivan, M. and Srite, M. (2005). Information technology and culture: Identifying fragmentary and holistic perspectives of culture. *Information and Organization*, 15(4):295–338.

Goffman, W. and Newill, V. A. (1964). Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204(4955):225–228.

Gurley, N. and Johnson, D. (2011). Viral Economics: An Epidemiological Model of Knowledge Diffusion in Economics. *Available at SSRN 1927126*.

Hamer, W. H. (1906). *Epidemic disease in England – the evidence of variability and the persistence of type*. Lancet.

Henrich, J. (2002). On modeling cognition and culture: Why cultural evolution does not require replication of representations. *Journal of Cognition and Culture*.

BIBLIOGRAPHY

Herbig, P. A. and Palumbo, F. (1994). The effect of culture on the adoption process: A comparison of Japanese and American behavior. *Technological Forecasting and Social Change*, 46(1):71–101.

Hethcote, H. W. (1976). Qualitative analyses of communicable disease models. *Mathematical Biosciences*, 28(3):335–356.

Hethcote, H. W. (1989). Three basic epidemiological models. pages 119–144.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and Control of Epidemics in a Globalized World. *Proceedings of the National Academy of Sciences*, 101:15124–15129.

Isaacs, S. (2001). The Power Distance between Users of Information Technology and Experts and Satisfaction with the Information System: Implication for Cross Cultural Transfer of IT. *Studies in health technology and informatics*, 84(2):1155–1157.

Kaba, B. and Osei-Bryson, K.-M. (2013). Examining influence of national culture on individuals' attitude and use of information and communication technology: Assessment of moderating effect of culture through cross countries study. *International Journal of Information Management*, 33(3):441–452.

Karmeshu and Pathria, R. K. (1980). Stochastic evolution of a nonlinear model of diffusion of information. *The Journal of Mathematical Sociology*, 7(1):59–71.

Kermack, W. O. and McKendrick, A. G. (1991). Contributions to the mathematical theory of epidemics—I. *Bulletin of Mathematical Biology*, 53(1):33–55.

Khine, M. S. and Saleh, I. M. (2011). *Models and Modeling: Cognitive Tools for Scientific Enquiry (Models and Modeling in Science Education)*. Springer, 1st edition. edition.

Kiss, I. Z., Broom, M., Craze, P. G., and Rafols, I. (2010). Can epidemic models describe the diffusion of topics across disciplines? *Journal of Informetrics*, 4(1):74–82.

Kleinbaum, D. G., Sullivan, K. M., and Barker, N. D. (2007). *A pocket guide to epidemiology*. Springer New York:.

Lambiotte, R. and Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190.

Last, J. M. and Abramson, J. H. (2001). A dictionary of epidemiology. 44.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2005). The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1).

Li, M. Y., Graef, J. R., Wang, L., and Karsai, J. (1999). Global dynamics of a SEIR model with varying total population size. *Mathematical Biosciences*, 160(2):191–213.

López-Pintado, D. (2008). Diffusion in complex social networks. *Games and Economic Behavior*, 62(2):573–590.

Lotka, A. J. (1922). The stability of the normal age distribution. *Proc. Nat. Acad. Sci. USA*, 8:339–345.

BIBLIOGRAPHY

Ma, Z. (2009). *Dynamical Modeling and Analysis of Epidemics*. World Scientific Publishing Company.

Mandal, S., Boslaugh, S., and Sinha, S. (2011). Mathematical models of malaria - a review. *Malaria Journal*, 10(1):202.

Marutschke, D. M. and Murao, H. (2013). Epidemiological Modeling of Knowledge Propagation in Scientific Publications. *ICIC Express Letters*, 7(3):923–928.

Marutschke, D. M., Nakajima, H., Tsuchiya, N., Yoneda, M., Iwami, T., and Kamei, K. (2009). Actualization of causality-based transparency and accuracy in system modeling with human-machine collaboration. *IC-MED Journal*, 3(2):131–141.

May, M. and Saitta, L. (2010). *Ubiquitous knowledge discovery: challenges, techniques, applications*, volume 6202. Springer.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467.

Mena-Lorcat, J. and Hethcote, H. W. (1992). Dynamic models of infectious diseases as regulators of population sizes. *Journal of Mathematical Biology*, 30(7):693–716.

M'Kendrick, A. G. (1925). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

Murillo, L. N., Murillo, M. S., and Perelson, A. S. (2013). Towards multiscale modeling of influenza infection. *Journal of Theoretical Biology*, 332(7):267–290.

Myatt, G. J. (2006). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. Wiley.

Newman, S. C. and Newman (2001). *Biostatistical Methods in Epidemiology*. Wiley-Interscience, 1 edition.

Nisbet, R., Elder, J., and Elder, J. (2009). Handbook of statistical analysis and data mining applications.

Okada, Y., Sakaki, T., Toriumi, T., Shinoda, K., Kazama, K., Noda, I., Numao, M., and Kuriha, S. (2013). 拡張 SIR モデルによる Twitter でのデマ拡散過程の解析 —False Rumor Diffusion Analysis based on The SIR-Extended Information Diffusion Model. *The th Annual Conference of the Japanese Society for Artificial Intelligence*,, pages 1–4.

Pearce, N. (2005). *A short introduction to epidemiology*. Centre for Public Health Research, Massey University.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996). Hiv-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586.

Perkins, R. and Neumayer, E. (2005). The international diffusion of new technologies: A multitechnology analysis of latecomer advantage and global economic integration. *Annals of the Association of American Geographers*, 95(4):789–808.

Pham, H. (2006). *Springer Handbook of Engineering Statistics*. Springer.

BIBLIOGRAPHY

Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., and Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising : Examining consumer responses and motivations to pass along email. *Journal of Advertising Research*, 44(4):333–348.

Pittel, B. (1987). On spreading a rumor. *SIAM Journal on Applied Mathematics*, 47(1):213–223.

Prematunge, C., Corace, K., McCarthy, A., Nair, R. C., Pugsley, R., and Garber, G. (2012). Factors influencing pandemic influenza vaccination of healthcare workers—A systematic review. *Vaccine*, 30(32):4733–4743.

Quinn, G. (2002). Experimental Design and Data Analysis for Biologists.

Reinhold Decker, H.-J. L. (2007). *Advances in Data Analysis*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer.

Rhodes, C. J., Jensen, H. J., and Anderson, R. M. (1997). On the critical behaviour of simple epidemics. *Proceedings of the Royal Society B Biological Sciences*, 264(1388):1639–1646.

Roberto, J., Piqueira, C., Navarro, B. F., Avenida, Gualberto, L., and Paulo, S. (2005). Epidemiological models applied to viruses in computer networks. *Journal of Computer Science*, 1(1):31–34.

Rohaya, D., Rambli, A., Matcha, W., Sulaiman, S., and Nayan, M. Y. (2012). Design and Development of an Interactive Augmented Reality Edutainment Storybook for Preschool. *IERI Procedia*, 2:802–807.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA. ACM.

Ross, R. (1911a). *The Prevention of Malaria*. Murray, London, 2nd edition.

Ross, R. (1911b). Some quantitative studies in epidemiology. *Nature*, 87(2188):466–467.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, third edition.

Salomon, J. A., Weinstein, M. C., Hammitt, J. K., , and Goldie, S. J. (2002). Empirically calibrated model of hepatitis c virus infection in the united states. *American Journal of Epidemiology*, 156(8):761–773.

Schuette, M. C. (2003). A qualitative analysis of a model for the transmission of varicella-zoster virus. *Mathematical Biosciences*, 182(2):113–126.

Shirai, T., Sakaki, S., Toriumi, F., Shinoda, K., Kazama, K., Noda, I., Numao, M., and Kurihara, S. (2012). Twitter ネットワークにおけるデマ拡散とデマ拡散防止モデルの推定 — Estimation of False Rumor Diffusion Model and Prevention Model of False Rumor Diffusion on Twitter. *Japanese Society of Artificial Intelligence*, pages 1–9.

Smith, R., Deitz, G., Royne, M. B., Hansen, J. D., Grünhagen, M., and Witte, C. (2013). Cross-cultural examination of online shopping behavior: A comparison of Norway, Germany, and the United States. *Journal of Business Research*, 66(3):328–335.

BIBLIOGRAPHY

Stauffer, D. and Sahimi, M. (2006). Discrete simulation of the dynamics of spread of extreme opinions in a society. *Physica A: Statistical Mechanics and its Applications*, 364:537–543.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Colizza, V., Isella, L., Régis, C., Pinton, J.-F., Khanafer, N., Van den Broeck, W., and Vanhems, P. (2011). Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 9(1):87.

Stewart, A. (2002). *Basic Statistics and Epidemiology - A practical guide*. Radcliffe Medical Press.

Sznajd-Weron, K. and Sznajd, J. (2001). Opinion evolution in closed community. *Arxiv preprint condmat0101130*, 11(6):13.

Tabah, A. N. (1999). Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology (ARIST))*, 34:249–286.

Taras, V., Steel, P., and Kirkman, B. L. (2012). Improving national cultural indices using a longitudinal meta-analysis of Hofstede's dimensions. *Journal of World Business*, 47(3):329–341.

Tchaicha, J. D. (2005). The Impact of Culture on Technology and Business: An Interdisciplinary, Experiential Course Paradigm. *Journal of Management Education*, 29(5):738–757.

Thompson, G. N., Estabrooks, C. A., and Degner, L. F. (2006). Clarifying the concepts in knowledge transfer: a literature review. *Journal of advanced nursing*, 53(6):691–701.

Van den Broeck, J. and Brestoff, J. R. (2013). *Epidemiology: Principles and Practical Guidelines*. Springer.

Wang, F., Zhang, Y., Wang, C., Ma, J., and Moon, S. (2010). Stability analysis of a SEIQV epidemic model for rapid spreading worms. *Computers & Security*, 29(4):410–418.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771.

Watts, D. J., Peretti, J., and Frumin, M. (2007). Viral marketing for the real world. *Harvard Business Review*, 85(5):22–23.

Witten, I. and Frank, E. (2011). Data mining: Practical machine learning tools and techniques.

Xiong, F., Liu, Y., Zhang, Z.-J., Zhu, J., and Zhang, Y. (2012). An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A*, 376(30-31):2103–2108.

Yip, S. and Rubia, T. D. (2010). *Scientific Modeling and Simulations (Lecture Notes in Computational Science and Engineering, 68)*. Springer.

Yuan, H. and Chen, G. (2008). Network virus-epidemic model with the point-to-group information propagation. *Applied Mathematics and Computation*, 206(1):357–367.

本論文に関連する研究発表

学術雑誌　論文発表　（査読あり）

1. D. Moritz Marutschke and Hajime Murao, "Parameter-Wise Clustering of Epidemiological Models to Map Knowledge Propagation in Scientific Publications," Innovations in Information and Communication Science and Technology (IICST 2012), Tomsk, Russia, Sept. 10-13, pp. 9-16, 2012

2. D. Moritz Marutschke and Hajime Murao, "Epidemiological Modeling of Knowledge Propagation in Scientific Publications," ICIC Express Letters, vol.7, no.3(B), pp. 923-928, 2013

3. D. Moritz Marutschke and Hajime Murao, "Cultural Characteristics of Knowledge Propagation in Scientific Publications – Japan, China, and Worldwide," Innovations in Information and Communication Science and Technology (IICST 2012), Tomsk, Russia, Sept. 2-5, 2013

4. D. Moritz Marutschke and Hajime Murao, "Short Study on Complexity and Feasibility of Deterministic Epidemiological Models to Track Knowledge Propagation in Scientific Publications," ICIC Express Letters, vol.8, no.4, pp. 1081-1088, 2014

国内会議　口頭発表　（査読無し）

1. D. Moritz Marutschke, Katsuari Kamei, and Hajime Murao, "Correlation of Heart Rate Variability and Stress related Subjective Variables," 第 24 回自律分散システムシンポジウム, Kobe, Japan, Jan. 27-28, 2012

2. D. Moritz Marutschke and Hajime Murao (マルチュケモリツ, 村尾元), "疫学モデルで科学論文内の情報伝播推定と知識クラスタリング," 第 18 回創発システムシンポジウム, Otsu, Japan, Sept. 1-3, 2012

3. D. Moritz Marutschke and Hajime Murao, "Refined Epidemiological Modeling of Knowledge Propagation in Scientific Publications with Hidden Knowledge Holders," 第 25 回自律分散システム・シンポジウム, Tohoku, Japan, Jan. 25-26, 2013

4. D. Moritz Marutschke and Hajime Murao (マルチュケ D. モリツ, 村尾元), "拡張した SEIRK 疫学的モデルで科学論文に基づく知識の伝搬トラッキング," SCI'13 第 57 回システム制御情報学会研究発表講演会, Hyogo, Japan, May 15-17, 2013

Appendices

APPENDIX A

NAMES AND SUBJECTS OF IEEE XPLORE DIGITAL LIBRARY

The names of publication details are as follows ("Aims and scope" as listed on each website see appendix B):

Acoustics, Speech and Signal Processing *Subject*: Signal Processing & Analysis

Control Theory and Applications *Subject*: Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Fields, Waves & Electromagnetics; Photonics & Electro-Optics; Robotics & Control Systems; Signal Processing & Analysis

Electronics & Communication Engineering Journal *Subject*: Communication, Networking & Broadcasting

Engineering & Technology *Subject*: Engineering Profession

Engineering Science and Education Journal *Subject*: Engineering Profession

IEEE Annals of the History of Computing *Subject*: Communication, Networking & Broadcasting; Computing & Processing (Hardware/Software); General Topics for Engineers (Math, Science & Engineering)

IEEE Computer *Subject*: Computing & Processing (Hardware/Software)

IEEE Computing & Control Engineering Journal *Subject*: Computing & Processing (Hardware/Software)

IEEE Control Systems *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profes-

sion; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Geoscience; Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications; Robotics& Control Systems; Signal Processing & Analysis; Transportation

IEEE Design & Test of Computers *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Geoscience; Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Network *Subject*: Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; General Topics for Engineers (Math, Science & Engineering); Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis

IEEE Software *Subject*: Computing & Processing (Hardware/Software)

IEEE Spectrum *Subject*: Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineering Profession; General Topics for Engineers (Math, Science & Engineering); Power, Energy, & Industry Applications

IEEE Technology and Society Magazine *Subject*: Engineering Profession; General Topics for Engineers (Math, Science & Engineering); Signal Processing & Analysis

IEEE Transactions on Automation Science and Engineering *Subject*: Robotics & Control Systems

IEEE Transactions on Communications *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Transactions on Energy Conversion *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves

& Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Nuclear Engineering; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Transactions on Evolutionary Computation *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Geoscience; Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Transactions on Fuzzy Systems *Subject*: Aerospace; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Power, Energy, & Industry Applications; Robotics & Control Systems; Signal Processing & Analysis; Transportation

IEEE Transactions on Image Processing *Subject*: Computing & Processing (Hardware/Software); General Topics for Engineers (Math, Science & Engineering); Signal Processing & Analysis

IEEE Transactions on Knowledge and Data Engineering *Subject*: Computing & Processing (Hardware/Software)

IEEE Transactions on Neural Networks *Subject*: Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Geoscience; Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis

IEEE Transactions on Pattern Analysis and Machine Intelligence *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; General Topics for Engineers (Math, Science & Engineering); Nuclear Engineering; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Transactions on Power Systems *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems;

Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Transactions on Signal Processing *Subject*: Signal Processing & Analysis

IEEE Transactions on Software Engineering *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Geoscience; Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications; Signal Processing & Analysis; Transportation

IEEE Transactions on Very Large Scale Integration (VLSI) Systems *Subject*: Aerospace; Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering); Geoscience; Nuclear Engineering; Photonics & Electro-Optics; Power, Energy, & Industry Applications;

Signal Processing & Analysis; Transportation

IEEE Transactions on Computers *Subject*: Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Photonics & Electro-Optics

Proceedings of the IRE and IEEE *Subject*: Engineering Profession; General Topics for Engineers (Math, Science & Engineering)

Transactions on Automatic Control *Subject*: Engineering Profession; General Topics for Engineers (Math, Science & Engineering); Signal Processing & Analysis

Transactions on Education *Subject*: Bioengineering; Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Engineered Materials, Dielectrics & Plasmas; Engineering Profession; Fields, Waves & Electromagnetics; General Topics for Engineers (Math, Science & Engineering)

Transactions on Information Theory *Subject*: Bioengineering; Computing & Processing (Hardware/Software)

APPENDIX B

AIMS AND SCOPE OF IEEE XPLORE DIGITAL LIBRARY

Control Theory and Applications *Aims and scope*: "IET Control Theory & Applications is devoted to control systems in the broadest sense, covering new theoretical results and the applications of new and established control methods. Among the topics of interest are system modeling, identification and simulation, the analysis and design of control systems (including computer-aided design), and practical implementation. The scope encompasses technological, economic, physiological (biomedical) and other systems, including man-machine interfaces. Most of the papers published deal with original work from industrial and government laboratories and universities, but subject reviews and tutorial expositions of current methods are welcomed. Correspondence discussing published

papers is also welcomed. Applications papers need not necessarily involve new theory; papers which describe new realizations of established methods, or control techniques applied in a novel situation, or practical studies which compare various designs, would be of interest. Of particular value are theoretical papers which discuss the applicability of new work, or applications which engender new theoretical applications."

Engineering & Technology *Aims and scope*: "Engineering & Technology is the IET's flagship magazine featuring analysis, news, innovation announcements, job advertisements and careers advice. The coverage is wide and aimed at professionals in all areas of engineering and technology including the key industry sectors of communications, control and automation, electronics, management, IT, manufacturing and power."

IEEE Annals of the History of Computing *Aims and scope*: "From the analytical engine to the supercomputer, from Pascal to von Neumann, from punched cards to CD-ROMs – the IEEE Annals of the History of Computing covers the breadth of computer history. Featuring scholarly articles by leading computer scientists and historians, as well as firsthand accounts by computer pioneers, the Annals is the primary publication for recording, analyzing, and debating the history of computing. The Annals also serves as a focal point for people interested in uncovering and preserving the records of this exciting field. The quarterly publication is an active center for the collection and dissemination of information on historical projects and organizations, oral history activities, and international conferences."

APPENDIX B. AIMS AND SCOPE OF IEEE XPLORE DIGITAL LIBRARY

IEEE Computer *Aims and scope*: "Computer, the flagship publication of the IEEE Computer Society, publishes highly acclaimed peer-reviewed articles written for and by professionals representing the full spectrum of computing technology from hardware to software and from current research to new applications. Providing more technical substance than trade magazines and more practical ideas than research journals. Computer delivers useful information that is applicable to everyday work environments."

IEEE Computing & Control Engineering Journal *Aims and scope*: "Computing & Control Engineering is aimed at practicing computing and control engineers. This magazine addresses the practice of computing software and information systems and their applications, together with control systems, automation and robotics, through articles on methods, techniques and processes currently used in industrial, commercial and business applications. Provides essential information to professional engineers of all disciplines currently engaged in the application of computer technology."

IEEE Control Systems *Aims and scope*: "IEEE Control Systems Magazine is the largest circulation technical periodical worldwide devoted to all aspects of control systems. The Magazine publishes tutorial and expository articles on all areas of control system design and applications. Authors are encouraged to submit articles on applications, design tools, control education, and applied research."

IEEE Design & Test of Computers *Aims and scope*: "IEEE Design & Test of Computers offers original works describing the methods used to design and test electronic product hardware and supportive software. The magazine focuses on current and

near-future practice, and includes tutorials, how-to articles, and real-world case studies. Topics include IC/module design, low-power design, electronic design automation, design/test verification, practical technology, and standards. IEEE Design & Test of Computers is cosponsored with the IEEE Council on Electronic Design and Automation, IEEE Circuits and Systems Society, and the IEEE Solid State Circuits Society."

IEEE Network *Aims and scope*: "IEEE Network was the number one most-cited journal in telecommunications, the number twelve most-cited journal in electrical and electronics engineering, and the number three most-cited journal in Computer Science Hardware and Architecture in 2004, according to the annual Journal Citation Report (2004 edition) published by the Institute for Scientific Information. This magazine covers topics which include: network protocols and architecture; protocol design and validation; communications software; network control, signaling and management; network implementation (LAN, MAN, WAN); and micro-to-host communications."

IEEE Software *Aims and scope*: "IEEE Software's mission is to build the community of leading and future software practitioners. The magazine delivers reliable, useful, leading-edge software development information to keep engineers and managers abreast of rapid technology change. The authority on translating software theory into practice, the magazine positions itself between pure research and pure practice, transferring ideas, methods, and experiences among researchers and engineers. Peer-reviewed articles, topical interviews, and columns by seasoned practitioners illuminate all aspects of the industry, including process improvement, project management, development tools, software main-

tenance, Web applications and opportunities, testing,usability, and much more."

IEEE Spectrum *Aims and scope*: "IEEE Spectrum Magazine, the flagship publication of the IEEE, explores the development, applications and implications of new technologies. It anticipates trends in engineering, science, and technology, and provides a forum for understanding, discussion and leadership in these areas. IEEE Spectrum is the world's leading engineering and scientific magazine. Read by over 300,000 engineers worldwide, Spectrum provides international coverage of all technical issues and advances in computers, communications, and electronics. Written in clear, concise language for the non-specialist, Spectrum's high editorial standards and worldwide resources ensure technical accuracy and state-of-the-art relevance."

IEEE Technology and Society Magazine *Aims and scope*: "The impact of technology (as embodied by the fields of interest in IEEE) on society, the impact of society on the engineering profession, the history of the societal aspects of electrotechnology, and professional, social, and economic responsibility in the practice of engineering and its related technology."

IEEE Transactions on Automation Science and Engineering *Aims and scope*: "T-ASE will publish foundational research on Automation: scientific methods and technologies that improve efficiency, productivity, quality, and reliability, specifically for methods, machines, and systems operating in structured environments over long periods, and the explicit structuring of environments. Its coverage will go beyond Automation's roots in mass production and include many new applications areas, such as Biotechnology,

pharmaceutical, and health care; Home, service, and retail; Construction, transportation, and security; Manufacturing, maintenance, and supply chains; and Food handling and processing. Research includes topics related to robots and intelligent machines/systems in structured environments and the explicit structuring of environments, and topics at the Operational/Enterprise levels such as System Modeling, Analysis, Performance Evaluation; Planning, Scheduling, Coordination; Risk Management; and Supply Chain Management. T-ASE will integrate knowledge across disciplines and industries."

IEEE Transactions on Communications *Aims and scope*: "IEEE Transactions on Communications was the number eight most cited journal in telecommunications in 2004, according to the annual Journal Citation Report (2004 edition) published by the Institute for Scientific Information. Read more at http://www.ieee.org/products/citations.html. This publication focuses on all telecommunications including telephone, telegraphy, facsimile, and point-to-point television, by electromagnetic propagation, including radio; wire; aerial, underground, coaxial, and submarine cables; waveguides, communication satellites, and lasers; in marine, aeronautical, space, and fixed station services; repeaters, radio relaying, signal storage, and regeneration; telecommunication error detection and correction; multiplexing and carrier techniques; communication switching systems; data communications; and communication theory."

IEEE Transactions on Energy Conversion *Aims and scope*: "The Transactions on Energy Conversion includes in its venue the analysis, control, planning, and economics of sources of electrical energy, distributed and cogeneration power plants, central station

grid connection, and equipment for generation and utilization of electric power, including electric machinery and energy storage systems."

IEEE Transactions on Evolutionary Computation *Aims and scope*: "Papers on application, design, and theory of evolutionary computation, with emphasis given to engineering systems and scientific applications. Evolutionary optimization, Machine Learning, intelligent systems design, image processing and machine vision, pattern recognition, evolutionary neurocomputing, evolutionary fuzzy systems, applications in biomedicine and biochemistry, robotics and control, mathematical modeling, civil, chemical, aeronautical, and industrial engineering applications."

IEEE Transactions on Fuzzy Systems *Aims and scope*: "The IEEE Transactions on Fuzzy Systems (TFS) is published bimonthly. TFS will consider papers that deal with the theory, design or an application of fuzzy systems ranging from hardware to software. Authors are encouraged to submit articles, which disclose significant technical achievements, exploratory developments, or performance studies of fielded systems based on fuzzy models. Emphasis will be given to engineering applications."

IEEE Transactions on Image Processing *Aims and scope*: "Signal-processing aspects of image processing, imaging systems, and image scanning, display, and printing. Includes theory, algorithms, and architectures for image coding, filtering, enhancement, restoration, segmentation, and motion estimation; image formation in tomography, radar, sonar, geophysics, astronomy, microscopy, and crystallography; image scanning, digital halftoning and display, and color reproduction."

IEEE Transactions on Knowledge and Data Engineering *Aims and scope*: "The IEEE Transactions on Knowledge and Data Engineering is an archival journal published monthly. The information published in this Transactions is designed to inform researchers, developers, managers, strategic planners, users, and others interested in state-of-the-art and state-of-the-practice activities in the knowledge and data engineering area. We are interested in well-defined theoretical results and empirical studies that have potential impact on the acquisition, management, storage, and graceful degeneration of knowledge and data, as well as in provision of knowledge and data services. Specific topics include, but are not limited to: a) artificial intelligence techniques, including speech, voice, graphics, images, and documents; b) knowledge and data engineering tools and techniques; c) parallel and distributed processing; d) real-time distributed; e) system architectures, integration, and modeling; f) database design, modeling and management; g) query design and implementation languages; h) distributed database control; j) algorithms for data and knowledge management; k) performance evaluation of algorithms and systems; l) data communications aspects; m) system applications and experience; n) knowledge-based and expert systems; and, o) integrity, security, and fault tolerance."

IEEE Transactions on Neural Networks *Aims and scope*: "IEEE Transactions on Neural Networks was the 7th most cited journal in electrical and electronics engineering in 2007, according to the annual Journal Citation Report (2007 edition), published by the Institute for Scientific Information. Read more at `http://www.ieee.org/products/citations.html`. Devoted to the science and technology of Neural Networks, which disclose significant technical knowledge, exploratory developments, and ap-

plications of Neural Networks from biology to software to hardware. Emphasis is on artificial Neural Networks."

IEEE Transactions on Pattern Analysis and Machine Intelligence *Aims and scope*: "The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) is published monthly. Its editorial board strives to present most important research results in areas within TPAMI's scope. This includes all traditional areas of computer vision and image understanding, all traditional areas of pattern analysis and recognition, and selected areas of machine intelligence. Areas of such Machine Learning, search techniques, document and handwriting analysis, medical image analysis, video and image sequence analysis, content-based retrieval of image and video, face and gesture recognition and relevant specialized hardware and/or software architectures are also covered."

IEEE Transactions on Power Systems *Aims and scope*: "Covers the requirements, planning, analysis, reliability, operation, and economics of electric generating, transmission, and distribution systems for general industrial, commercial, public, and domestic consumption."

IEEE Transactions on Signal Processing *Aims and scope*: "The IEEE Transactions on Signal Processing covers novel theory, algorithms, performance analyses and applications of techniques for the processing, understanding, learning, retrieval, mining, and extraction of information from signals. The term "signal" includes, among others, audio, video, speech, image, communication, geophysical, sonar, radar, medical and musical signals. Examples of topics of interest include, but are not limited to, information process-

ing and the theory and application of filtering, coding, transmitting, estimating, detecting, analyzing, recognizing, synthesizing, recording, and reproducing signals."

IEEE Transactions on Software Engineering *Aims and scope*: "The IEEE Transactions on Software Engineering is an archival journal published monthly. We are interested in well-defined theoretical results and empirical studies that have potential impact on the construction, analysis, or management of software. The scope of this Transactions ranges from the mechanisms through the development of principles to the application of those principles to specific environments. Since the journal is archival, it is assumed that the ideas presented are important, have been well analyzed, and/or empirically validated and are of value to the software engineering research or practitioner community. Specific topic areas include: a) development and maintenance methods and models, e.g., techniques and principles for the specification, design, and implementation of software systems, including notations and process models; b) assessment methods, e.g., software tests and validation, reliability models, test and diagnosis procedures, software redundancy and design for error control, and the measurements and evaluation of various aspects of the process and product; c) software project management, e.g., productivity factors, cost models, schedule and organizational issues, standards; d) tools and environments, e.g., specific tools, integrated tool environments including the associated architectures, databases, and parallel and distributed processing issues; e) system issues, e.g., hardware-software trade-off; and f) state-of-the-art surveys that provide a synthesis and comprehensive review of the historical development of one particular area of interest."

APPENDIX B. AIMS AND SCOPE OF IEEE XPLORE DIGITAL LIBRARY

IEEE Transactions on Very Large Scale Integration (VLSI) Systems

Aims and scope: "Includes all major aspects of the design and implementation of VLSI/ULSI and microelectronic systems. Topics of special interest include: systems specifications, design and partitioning, high performance computing and communication systems, Neural Networks, wafer-scale integration and multichip module systems and their applications."

IEEE Transactions on Computers *Aims and scope*: "IEEE Transactions on Computers was the number sixteen most cited journal in electrical and electronics engineering in 2004 according to the annual Journal Citation Report (2004 edition), published by the Institute for Scientific Information. Read more at `http://www.ieee.org/products/citations.html`. The IEEE Transactions on Computers is a monthly publication with a wide distribution to researchers, developers, technical managers, and educators in the computer field. It publishes papers, brief contributions, and comments on research in areas of current interest to the readers. These areas include, but are not limited to, the following: a) computer organizations and architectures; b) operating systems, software systems, and communication protocols; c) real-time systems and embedded systems; d) digital devices, computer components, and interconnection networks; e) specification, design, prototyping, and testing methods and tools; f) performance, fault tolerance, reliability, security, and testability; g) case studies and experimental and theoretical evaluations; and h) new and important applications and trends."

Proceedings of the IRE and IEEE *Aims and scope*: "The most highly-cited general interest journal in electrical engineering and computer science, the Proceedings is the

best way to stay informed on an exemplary range of topics. This journal also holds the distinction of having the longest useful archival life of any EE or computer related journal in the world! Since 1913, the Proceedings of the IEEE has been the leading journal to provide in-depth tutorial and review coverage of the technical developments that shape our world."

Transactions on Automatic Control *Aims and scope*: "In the IEEE Transactions on Automatic Control, the IEEE Control Systems Society publishes high-quality papers on the theory, design, and applications of control engineering. Two types of contributions are regularly considered: 1) Papers: Presentation of significant research, development, or application of control concepts. 2) Technical Notes and Correspondence: Brief technical notes, comments on published areas or established control topics, corrections to papers and notes published in the Transactions. In addition, special papers (tutorials, surveys, and perspectives on the theory and applications of control systems topics) are solicited."

Transactions on Education *Aims and scope*: "The aims of the IEEE Transactions on Education are both scientific and educational, grounded in the theory and practice of electrical and computer engineering. The scope covers education research, methods, materials, programs, and technology in electrical engineering, computer engineering, and fields within the scope of interest of IEEE. Manuscripts submitted to the Transactions should clearly embrace one or more of these topic areas."

Transactions on Information Theory *Aims and scope*: "The IEEE TRANSACTIONS ON INFORMATION THEORY publishes papers concerned with the transmission,

processing, and utilization of information. While the boundaries of acceptable subject matter are intentionally not sharply delimited, its scope currently includes Shannon theory, coding theory and techniques, data compression, sequences, signal processing, detection and estimation, pattern recognition, learning and inference, communications and communication networks, complexity and cryptography, and quantum information theory and coding. IEEE Transactions on Information Theory papers normally contain a strong conceptual and/or analytical contribution."