PAI

► Positional Accuracy Improvement (PAI)

Pandemics, Detection and Management

CHRISTOPHER L. BARRETT, STEPHEN EUBANK, BRYAN LEWIS, MADHAV V. MARATHE
Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

Synonyms

Epidemiology, computational; Epidemics; Public health; Information management systems; Distributed information systems

Definition

Epidemiology is the study of patterns of health in a population and the factors that contribute to these patterns. It plays an essential role in public health through the elucidation of the processes that lead to ill health as well as the evaluation of strategies designed to promote good health. Epidemiologists are primarily concerned with public health data, which includes the design of studies, evaluation and interpretation of public health data, and the maintenance of data collection systems.

Computational Epidemiology is the development and use of computer models for the spatio-temporal diffusion of disease through populations. The models may range from descriptive, e. g. static estimates of correlations within large databases, to generative, e. g. computing the spread of disease via person-to-person interactions through a large population. The disease may represent an actual infectious disease, or it may represent a more general reaction-diffusion process, such as the diffusion of innovation. The populations of interest depend on the disease, including humans, animals, plants, and computers. Similarly, the interactions that must be represented depend on the disease and the populations, including physical proximity for aerosol-borne disease, sexual contact for sexually transmitted diseases, and insect feeding patterns for mosquito-borne diseases. In general, then, computational epidemiology creates computer models of diffusive processes spreading across interaction networks.

The basic goal of epidemiological modeling is to understand the dynamics of disease spread well enough to control it. Potential interventions for controlling infectious disease include pharmaceuticals for treatment or prophylaxis, social interventions designed to change transmission rates between individuals, physical barriers to transmission, and eradication of vectors. Efficient use of these interventions requires targeting subpopulations that are on the critical path of disease spread. Computational models can be used to identify those critical subpopulations and to assess the feasibility and effectiveness of proposed interventions.

Historical Background

Epidemiology did not emerge as a distinct discipline until the mid-19th century as the medical sciences sought to determine the efficacy of different medical practices. John Snow famously interrupted the 1854 cholera outbreak in London by removing the handle of the Broad Street pump, an event that is widely credited with bringing epidemiology into the mainstream. His studies along with those of many others were responsible for bringing about wide-ranging public health reforms and laid the foundation for the development of the germ theory of disease causation. Once etiological agents were identified as the cause of disease, the sanitary reforms of the late 19th and early 20th centuries greatly reduced the incidence of infectious disease in the human population. Epidemiology continued to identify novel causes of disease but also matured and began to consider the social determinants of health. Aided by improved statistical tools epidemiologists were able to focus on more nuanced analysis of population-wide health data, for example, linking lung cancer to smoking. The pharmaceutical revolution of the mid-20th century required epidemiology to assess the efficacies of the new

therapies being created. Epidemiology has further gained from the ability to manipulate large bodies of data and perform complex calculations on this data using computers. As technology has improved the sophistication of techniques to analyze public health data used by epidemiologists has kept pace. Recent years have seen the emergence of computational epidemiology which focuses on using complex computer models and innovative forms of data analysis. At its core, epidemiology is still focused on improving the health of the public, however, recently, epidemiologically inspired methods and techniques have been harnessed to analyze computer security, network routing, distributed databases, marketing strategies, and other social phenomena.

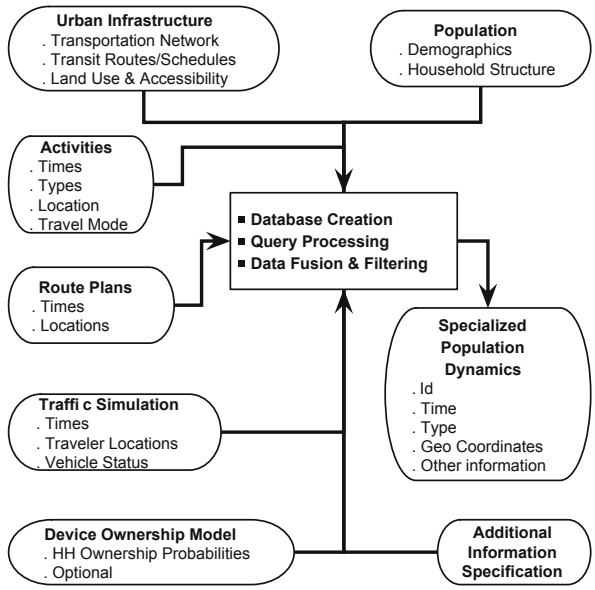
Scientific Fundamentals

The spread of infectious disease depends both on properties of the pathogen and the host. An important factor that greatly influences an outbreak of an infectious disease is the structure of the interaction network across which it spreads. Descriptive models are useful for estimating properties of the disease, but the structure of the interaction network changes with time and is often affected by the presence of disease and public health interventions. Thus generative models are most often used to study the effects of public health policies on the control of disease.

Aggregate or collective computational epidemiology models often assume a population is partitioned into a few subpopulations (e. g. by age) with a regular interaction structure within and between subpopulations. The resulting model can typically be expressed as a set of coupled ordinary differential equations. Such models focus on estimating the number of infected individuals as a function of time, and have been useful in understanding population-wide interventions. For example, they can be used to determine the level of immunization required to create herd immunity.

Disaggregate or individual-based models, in contrast, represent each interaction between individuals, and can thus be used to study critical pathways. Disaggregate models require neither partitions of the population nor assumptions about large scale regularity of interactions; instead, they require detailed estimates of transmissibility between individuals. The resulting model is typically a stochastic finite discrete dynamical system. For more than a few individuals, the state space of possible configurations of the dynamical system is so large that they are best studied using computer simulation.

GIS tools and techniques play an important role in building these computational tools. The overall approach followed by disaggregate models consists of the following steps:



Pandemics, Detection and Management, Figure 1 Schematic diagram showing how various databases and integrated to create a synthetic population. GIS plays an important role in constructing these synthetic populations

Step 1 Creating a set of (agent) synthetic interactors,
Step 2 Generating (time varying) interaction networks,
Step 3 Detailed simulation of the epidemic process.

Step 1 creates a synthetic urban populations [3,4,5,6], and is done by integrating a variety of databases from commercial and public sources into a common architecture for data exchange that preserves the confidentiality of the original data sets, yet produces realistic attributes and demographics for the synthetic individuals. Figure 1 shows a schematic diagram. The synthetic population is a set of synthetic people, each associated with demographic variables drawn from any of the demographics available in the census [3,7]. Joint demographic distributions can be reconstructed from the marginal distributions available in typical census data using an iterative proportional fitting (IPF) technique. Each synthetic individual is placed in a household with other synthetic people and each household is located geographically in such a way that a census of our synthetic population yields results that are statistically indistinguishable from the original census data, if they are both aggregated to the block group level. Synthetic populations are thus statistically indistinguishable from the census data; nevertheless, since they are synthetic they respect privacy of individuals within the population. Note that, census tables are precisely constructed so as to respect privacy. The *synthetic individuals* carry with them a complete range of demographic attributes collected in the census data. This include variables such as income level, age, etc.

In Step 2, a set of activity templates for households are determined based on several thousand responses to an activity or time-use survey. These activity templates include the sort of activities each household member performs and the time of day they are performed. Each synthetic household is matched with one of the survey households, using a decision tree based on demographics such as the number of workers in the household, number of children of various ages, etc. The synthetic household is assigned the activity template of its matching survey household. For each household and each activity performed by this household, a preliminary assignment of a location is made based on observed land-use patterns, tax data, etc. This guess must be calibrated against observed travel-time distributions. However, the travel-times corresponding to any particular assignment of activities to locations cannot be determined analytically. Using sophisticated techniques in combinatorial optimization, machine learning and agent based modeling the populations, their activity locations and their itineraries [3,5] are refined so as to be structurally and statistically consistent. See Fig. 1 for a schematic diagram. Thus for a city – demographic information for each person and location, and a minute-by-minute schedule of each person's activities and the locations where these activities take place is generated by a combination of simulation and data fusion techniques. This forms the basis of the interaction network that can be abstractly represented by a (vertex and edge) labeled bipartite graph G_{PL}, where P is the set of people and L is the set of locations. If a person $p \in P$ visits a location $l \in L$, there is an edge $(p, l, \text{label}) \in E(G_{PL})$ between them, where label is a record of the type of activity of the visit and its start and end points. Each vertex (person and location) can also have labels. The person labels correspond to his/her demographic attributes such as age, income, etc. The labels attached to locations specify the location's attributes such as its x and y coordinates, the type of activity performed, maximum capacity, etc. Note that, there can be multiple edges between a person and a location recording different visits.

Step 3 consists of developing computational model for representing the disease within individual interactor and its transmission between interactors. The model can be viewed as a *coupled probabilistic timed finite state machine*. Each individual is associated with a timed probabilistic finite state machine – the state transitions are probabilistic; the transitions may be timed – i. e. they may occur at a specified time after the previous transition – or there may be a fixed probability of transition for each discrete time interval. Furthermore, the automata are coupled to other automata – this coupling is derived from the social contact network. The state of the automata corresponding to an individual are updated probabilistically based on the current state of the individual and the disease state of his neighbors. This state transition is probabilistic and depends on the duration of contact. It may also depend on the attributes of the people involved (age, profession, health status, etc.) as well as the type of contact (intimate, casual, etc.), and it might not be symmetric (a child is more likely to infect a teacher than the other way around). Again GIS tools play an important role in constructing the models. The tools include: methods for integrating spatio-temporal surveillance data, mapping of the disease outbreak to geographic locations, designing intervention effective strategies, such as closing specific public locations, social distancing, etc.

An integrated information management system can now be constructed using these computational models. The systems is usually event triggered and the supporting system should ideally self organize in response to such a trigger. Figure 2 shows a possible architecture for such a system.

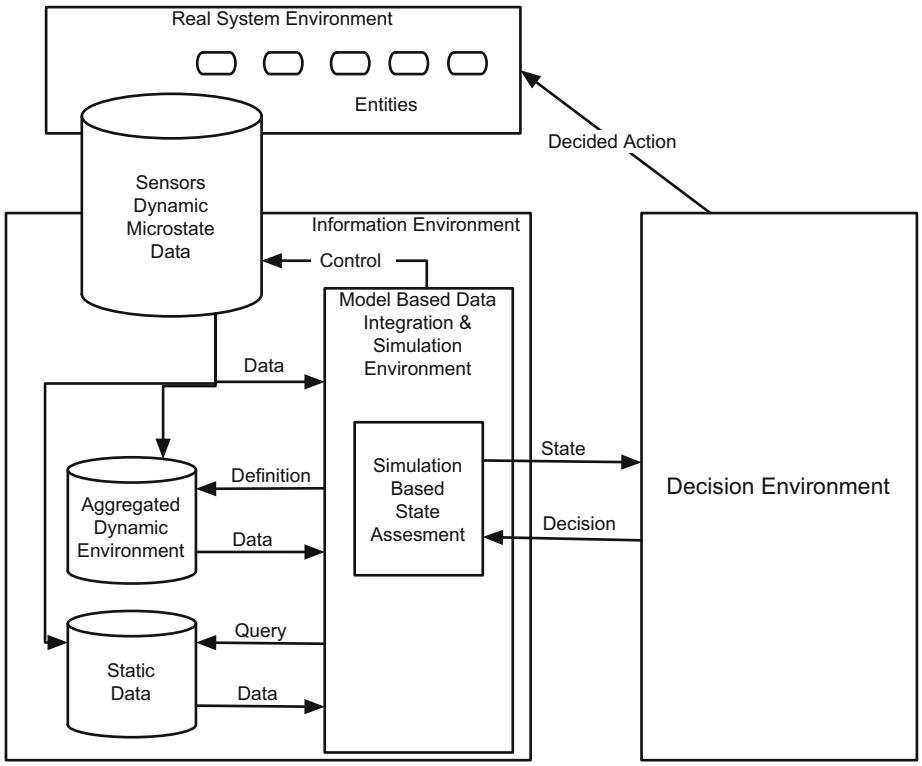
Application Areas

Epidemiological simulations are a subclass of more general interaction processes called reaction-diffusion processes. In their general form, such systems consist of a set of entities (interactors) and an interaction-hypergraph denoting neighborhood relationships among interactors. At each time step, based on certain criteria, subsets of neighboring interactors interact. The interactions can result in two things: (i) the state or the function of the interactors can change, the interaction-hypergraph can change. A number of applications in physical and social sciences that can be viewed in this framework. These include: (i) physical systems such as n-body dynamics, (ii) biological systems such as bio-chemical reactions, (iii) social systems such as diffusion of norms and fads, (iv) public health systems such as epidemics, (v) communication networks such as spread of worms on the Internet, routing of packets and updating distributed databases, (vi) business and information systems such as viral marketing, etc.

The systems differ in the relative rates at which interactions happen as compared to the change in the state of the interactors, and the network structure. For example, in epidemics, an individual when exposed to an infectious disease can become infected after a certain time period that depends on the demographic properties of the individual and the disease characteristics. When routing packets over a wireless ad-hoc network, the state of the interactor and the the connectivity of the underlying network changes rapidly.

Public Health

Computational epidemiology is a new and fast growing branch of public health. Using sophisticated and highly



Pandemics, Detection and Management, Figure 2
Schematic diagram for constructing a simulation based integrated information management system

tuned computer simulations, different public health interventions can be evaluated that would be unfeasible and/or unethical in the real world. Furthermore, as these techniques become more and more sophisticated, these population and disease dynamics can be better analyzed *in silico* than *in vivo*. See [6,8,12,13,14,15,17,18,23].

Social Sciences

Social Sciences have a rich history of studying social phenomenon using epidemic style models. This includes, diffusion of norms, fads, etc [21,26,29].

Epidemic/Viral Marketing and Advertising

This class of applications consist of marketing techniques to spread brand awareness. The information about brands, products, etc. can be exchanged via word of mouth using social networks that capture individuals meeting each other in physical or cyberspace (via e. g. blogs, chat rooms). Viral marketing is popular because it is usually easy and affordable to execute the marketing campaign. It is becoming all the more popular due to the Internet which allows for much more rapid dissemination.

Computer Network Security

Epidemic style models are being used to study the spread of worms and viruses on the Internet [2,27]. Computer models can be used to study the propagation of viruses as well as ways to control its spread. A unique feature of these systems is that unlike biological systems, computer viruses usually spread extremely fast, usually in a matter of hours if not minutes. Moreover, the viruses are synthetic; humans construct these viruses and their genetic variations.

Distributed Computing, Communication and Information Systems

A number of tasks in computing, communication and information systems can be achieved by using epidemic style algorithms. This includes: routing using local information under unreliable conditions [10], location of resources [11], and updating replicated distributed databases [9,11].

Acknowledgement

The work was supported in part by NIH MIDAS project.

Cross References

- ► Bioinformatics, Spatial Aspects
- ► Biomedical Data Mining, Spatial
- ► Data Analysis, Spatial
- ► Exploratory Spatial Analysis in Disease Ecology

Recommended Reading

1. Albert, R., Barabasi, A.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
2. Boguna, M., Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Complex Networks with Degree Correlations. In: Pastor-Satorras, R., Rubi, M., Diaz-Guilera, A. (eds.) Statistical Mechanics of Complex Networks. Lect. Notes Phys. **625**, 127–147 (2003)
3. Barrett, C., Beckman, R., Berkbigler, K., Bisset, K., Bush, B., Campbell, K., Eubank, S., Henson, K., Hurford, J., Kubicek, D., Marathe, M., Romero, P., Smith, J., Smith, L., Speckman, P., Stretz, P., Thayer, G., Eeckhout, E., Williams, M.D.: TRANSIMS: Transportation Analysis Simulation System. Technical Report LA-UR-00–1725, Los Alamos National Laboratory Unclassified Report. An earlier version appears as a 7 part technical report series LA-UR-99-1658 and LA-UR-99-2574 to LA-UR-99–2580 (2001)
4. Beckman, R.J., et al: TRANSIMS-Release 1.0: The Dallas-Fort Worth Case Study, Technical Report LA-UR-97-4502, Los Alamos National Laboratory (1997)
5. Barrett, C., Eubank, S., Marathe, M.: Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction Based Approach, to appear. In: Goldin, D., Smolka, S., Wegner, P. (eds.) Interactive Computation: The New Paradigm. Springer, New York (2005)
6. Barrett, C., Smith, J.P., Eubank, S.: Modern Epidemiology Modeling. Sci. Am. **292**(3), 54–61 (2005)
7. Beckman, R., Baggerly, K., McKay, M.: Creating synthetic baseline populations. Transportation Research Part A, Policy Pract. **30**, 415–429 (1996)
8. Carley, K., Fridsma, D., Casman, E., Altman, N., Chang, J., J., Kaminski, B., Nave, D., Yahja, A.A.: BioWar: Scalable Multi-Agent Social and Epidemiological Simulation of Bioterrorism Events. NAACSOS conference proceedings, Pittsburgh, PA (2003)
9. Kempe, D., Kleinberg, J.M., Demers, A.J.: Spatial gossip and resource location protocols. J. ACM **51**(6), 943–967 (2004)
10. Intanagonwiwat, C., Govindan, R., Estrin, D., Heidemann, J., Silva, F.: Directed Diffusion for Wireless Sensor Networking. ACM/IEEE Trans. Netw. **11**(1), 2–16 (2002)
11. Demers, A.J., Greene, D.H., Hauser, C., Irish, W., Larson, J., Shenker, S., Sturgis, H.E., Swinehart, D.C., Terry, D.B.: Epidemic Algorithms for Replicated Database Maintenance. PODC 1–12 (1987)
12. Eubank, S., Guclu, H., Anil Kumar, V.S., Marathe, M., Srinivasan, A., Toroczkai, Z., Wang, N.: Modeling Disease Outbreaks in Realistic Urban Social Networks. Nature **429**, 180–184 (2004)
13. Eubank, S., Anil Kumar, V.S., Marathe, M., Srinivasan, A., Wang, N.: Structure of Social Contact Networks and their Impact on Epidemics. In: AMS-DIMACS Special Volume on Epidemiology **70**, 181–213 (2005)
14. Ferguson, N.L., Cummings, D.A.T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Lamsirithaworn, S., Burke, D.S.: Strategies for containing an emerging influenza pandemic in Southeast Asia, Nature **437**, 209–214 (2005)
15. Ferguson, N.L., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S.: Strategies for mitigating an influenza pandemic, Nature, 448–452 (2006)
16. Ganesh, A., Massoulie, L., Towsley, D.: The Effect of Network Topology on the Spread of Epidemics. *IEEE Infocom* 2005, Miami, FL (2005)
17. Germann, T.C., Kadau, K., Longini Jr., I.M., Macken, C.A.: Mitigation strategies for pandemic influenza in the United States. Proc. of National Academy of Sciences (PNAS), vol. 103(15), 5935–5940 (2006)
18. Hethcote, H.: The Mathematics of Infectious Diseases. SIAM Rev. **42**(4), 599–653 (2000)
19. Jurvetson, S.: What exactly is viral marketing? Red Herring **78**, 110–112 (2000)
20. Kephart, J.O., White, S.R.: Directed-graph epidemiological models of computer viruses. Proc. 1991 IEEE Computer Society Symposium on Research in Security and Privacy, pp. 343–359 (1991)
21. Macy, M., Willer, R.: From Factors to Actors: Computational Sociology and Agent-Based Modeling. Ann. Rev. Soc. **28**, 143–166 (2002)
22. Monge, P., Contractor, N.: Theories of Communication Networks. Oxford University Press, New York (2003)
23. Meyers, L., Newman, M.E.J., Martin, M., Schrag, S.: Applying network theory to epidemics: Control measures for outbreaks of mycoplasma pneumonia. Emerging Infectious Diseases**9**(2), 204–210 (2003)
24. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: WWW '004 (2004)
25. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
26. Rogers, E.M.: Diffusion of Innovations, 4th edn. Free Press, New York (1995)
27. Savage, S., Voelker, G.M., Paxson, G.V.V., Weaver, N.: Center for Internet Epidemiology and Defenses. http://www.ccied.org/
28. Schelling, T.: Micromotives and Macrobehavior, W.W., Norton (1978)
29. Vega-Redondo, F.: Diffusion, Search and Play in Complex Social Networks, forthcoming. Econometric Society Monograph Series (2006)
30. Lilienfeld, D.E.: Foundations of Epidemiology, 3rd edn. Oxford University Press, Oxford (1994)

Parallel Computing

▶ Distributed Geospatial Computing (DGC)

Parametric Model

▶ Hurricane Wind Fields, Multivariate Modeling

Partial Order

▶ Hierarchies and Level of Detail

Participation Index

▶ Co-location Patterns, Algorithms

Participation Ratio

▶ Co-location Patterns, Algorithms

Participatory Planning and GIS

CHRISTINE ROTTENBACHER
Institute of Geoinformation and Cartography,
TU Vienna, Vienna, Austria

Synonyms

Moved planning process; Negotiation; Spatial decision making of groups; Sincerity; Walking, joint; Self-referential context; Group decisions

Definition

In participatory planning processes planners have the task to organize the process of negotiation in a field of different interests and to develop a common space for acting. In participative planning processes planners care for an understanding of all steps of informing and decision making. Planners aim for sincerity, avoiding manipulation and incorrectness, showing cost, benefits and risks of decision [13,26].

Historical Background

Building a planning process from the human resources (skills, knowledge, ideas) of the participants contributes to the development of self responsibility and self initiative of the participating group. The assumption is that the shared possibilities of understanding, decision making, and acting are growing and enable a planning process of balance between an orientation-giving concept and a patchwork of decisions and actions. This orienting concept corresponds with the spatial developments which can be solved within a concrete task through an orientation in values [11,12].

Planners have to organize participative planning processes to develop a common goal in space. A participative planning process occurs in a shared constructed reality related to the concrete space. Shared reality emerges out of *interaction, understanding, decision making, negotiation, realizing, and the shared experiencing of the outcome* within groups [1,2,7,23].

The *moved planning process* is a special design of planning participation processes for empowerment of participants [21]. Usually the participation processes last for two or more years. The design of these processes is based on interdisciplinary research results about knowledge construction and decision making in groups. The research covers the changing interaction structure of the attentive subgroups of participants, the building of networks, and the decision making process of the groups.

The mutual experience of *concrete situations* create corresponding structures between participants [20]. Groups of people affected by a plan are invited to show their daily environment, the "object" of the plan. Every meeting is started with walking together through the space to be planned. After the walk follows a sitting period together to reflect the shared experiences, to find topics and tasks until the next meeting. During the walk starts an exchange about the experience and knowledge about the special place; everybody is an expert of her daily live with the place [8].

- During concrete experiences the meanings get related to the ever changing shared experiences. The previously constructed reality of individual participants often differs from the encountered reality during the walk. Differences can be pointed out and erroneous conceptions corrected. For example, plants that indicate the presence of nutrients show directly a fertilizer usage. The feedback from the visible evidence forces the participants to learn about the consequences of their actions. In the shared experience and speaking about it different realities are explained.
- The participants get familiar with the walking situation, find their style of interaction and take over roles as experts of their daily environment and show their meanings and usage of space. Then they experience movement and space.
- Body exercises, which require coordination of different body parts, strengthen the connection of neurons between many parts of the central nervous system. Even simple walking activates different neuronal networks for motor control, integrate vestibular systems, and optical and acoustical stimuli. Simple walking leads further to a parallel processing for integration of perception. Walking regularly creates an awareness of perceiving, feeling, thinking, an acting [10].
- Movement is orientated action. Joint movement leads to actions together. The theory behind different body therapy forms deals with the connection of involuntary movement and acting, and leads to insights about movement together and common acting [15]. To each personal movement-behavior belongs a shared movement-behavior [18]. This effect is exploited in groups which work with body therapeutic methods, e. g. as in authentic movement [19].

Joint walking creates experiences together, namely the common rhythm of step, the regular breathing, the physical effort, and the feeling of fatigue. Pictures are perceived through senses and attending people are encouraged to trust in their senses while moving their bodies.

Planning as an instrument for preparing decisions depends on various factors. It happens in a social system of different pressure groups. There are advices, rules, administrative and technical restrictions, right in ownership and neighborhood. There are political influences or competing projects [16].

Decision making, acting, and bearing responsibility are connected processes within the planning group. A planning group consists of person responsible, planners, and of people affected by a planning intention; at the beginning of the planning process, they form a heterogeneous group. There is no shared language, no shared way of looking at issues, no common assumptions. Every participant brings her histories and perspectives to the group. From the beginning it is necessary to look at differences and commonalities. In order to transform needs and interests into a more comprehensive understanding (which takes other needs into account), it is necessary to look at the process of shared knowledge construction in groups. Participants cannot transcend their particularity [11]. If participants make decisions appropriate to their personal context, they have to express (and need to get space for doing so) their particularity to others, and learn about the particularities of the other. This leads to a shared knowledge based on concrete situations [14]. Participants have particular knowledge that arises from experiences, also experiences in their social positions, and those social positioning influence the assumptions and interests they bring to the meeting.

Scientific Fundamentals

The Human Theory of Action [9] describes a human as an organism-environment entity. A human is embedded as a body-mind person in a social ecological environment. Experiences and knowledge are stored up in the body, and used in each concrete situation [18]. Related to these concrete situations are the possibilities of reflection and imagination. According to this theory knowledge construction and action are immediately connected.

A human is active, orientated in the future [9]. She puts her own theories and goals, and makes hypothesis about the outcome of her acting in everyday acting. These hypotheses are immediately verified through acting, and corrections of the actions. Integrated in a planning theory, it is important to consider how to integrate steps of decision making, experiencing the outcome of these decisions into the technical issues and the design of a planning process [26]. Decisions are made continuously during the planning steps.

Mutual understanding between people is possible, if the coding and decoding participants have corresponding perceiving capacities and interpretation patterns. When a planner comes to a planning group of a village a base for shared perceiving and interpretation patterns has to be found. Understanding of perceived acts becomes meaningful by the different context of social interactions.

The personal acting has a self referential meaning to give continuity and identity to the own being and acting. The "*I*" is a live system which has emotional, spiritual and cognitive abilities. This is adapting permanently the changing environment. For the internal structure it is crucial to have self awareness for these changes and anchor them to identity. A personal history is experienced and humans develop through those experiences. New competences are acquired through the contact with others.

From the base of this "*I*" it is possible to get into contact with "*you*" and with the experiences of the "*you*" [6]. Within this self referential context it is easier to understand behavior of the others as an expression of their "*I*".

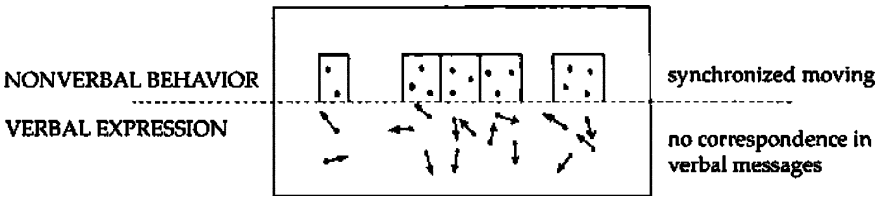
At a meeting and a joint walk this concrete context is strengthened. People find a common rhythm and breath, and have joint experiences. The immediate behavior evokes primarily patterns of existence and less patterns of thinking. New experiences, understanding, and knowledge are shared [14]. The participants experience themselves mutually within every new meeting and the structure of interaction in the concrete situation is new defined.

Participants develop new roles and test them in new behaving patterns and acting. This is the condition that they are able to perceive new contents and information, and integrate it into their personal experience, knowledge, and acting. This flexible interaction enables to anticipate and imagine the future [17].

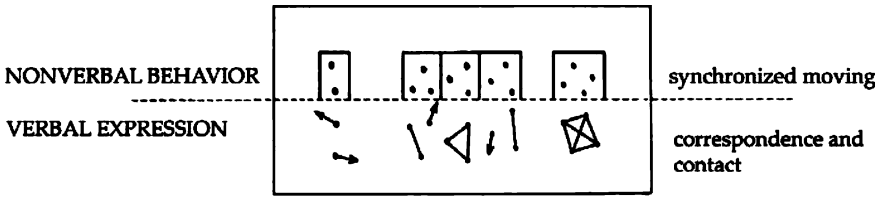
During walking the meaning of symbols are related to the ever changing joint experiences. Our understandings of symbols and behavior are exchanged. Not only are meanings expressed, but humans become more aware of what others think of them [8]. Through communication, humans look at themselves through the perspective of their partners, and take on their role. Mutual expectations have an effect on communication. When a person is walking, the location is present as a context for talking. There is one base of experiencing an understanding reality.

The research points out that concrete experienced correspondence in a group links participants and leads to grounded knowledge construction and decision making. Empirical evidence shows that networks between participants are strengthened because of the common experience in concrete situations [14,15,18]. The usual group-, and decision structures are enlarged and new ones rebuilt by experiencing movement and living space; and a shared reality grows avoiding a bargaining of interests. Walking in concrete situations contributes to the decision process in groups because it strengthens the concrete experience.

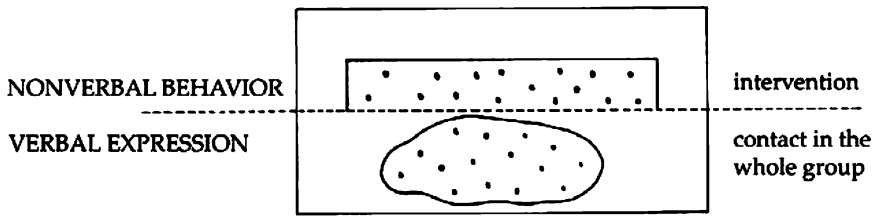
Participants form subgroups reflecting the social structure of the village. Walking breaks up subgroups and aids interactions among all participants [3]. The emotional correspondence is increasing in these changing subgroups. After the start of walking the different subgroups reach



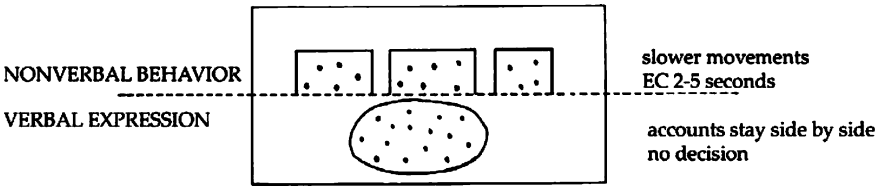
Participatory Planning and GIS, Figure 1 Correspondence in Movement



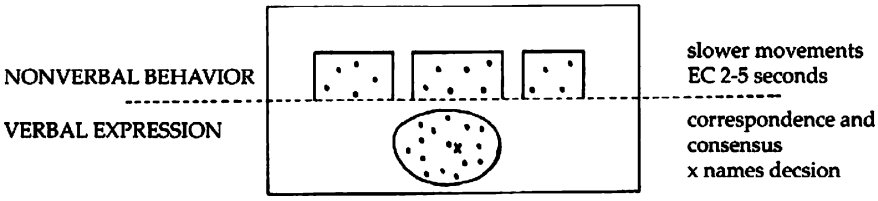
Participatory Planning and GIS, Figure 2 Correspondence in Movement and Contact in Verbal Expressions



Participatory Planning and GIS, Figure 3 Correspondence and Contact in the Whole Group



Participatory Planning and GIS, Figure 4 Correspondence without Decision in the Whole Group



Participatory Planning and GIS, Figure 5 Correspondence with Decision in the Whole Group

a synchronized moving which is observable [4,5]. In comparison with verbal expressions it is ascertainable that at that time there was mostly no correspondence in verbal expressions as shown in Fig. 1.

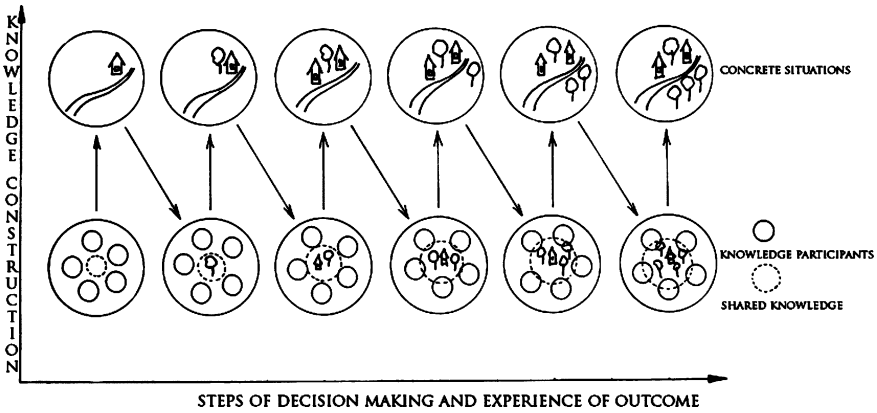
Figure 2 shows the beginning verbal correspondence and contact in the subgroups.

Figure 3 shows an event from the whole group. The subgroups open and experience together a concrete situation, exchange state, attitudes, and meanings. After this "intervention" the participants keep together, and build new subgroups within the whole group relating to the topics they found together.

Figures 4 and 5 show generalizations of events during the sitting period within the whole group. An emotional correspondence in nonverbal behavior is observable. Typically the movements get slower, and participants turn towards each other. Figure 4 describes the situation when participants stay side by side without discussion and without a decision.

Figure 5 describes the comparison of nonverbal correspondence with verbal expression. Here a person named a decision and all participants agreed. These decisions are then realized quickly.

During the sitting phase participants narrate in the first person. They start "naming" what they experienced and perceived, and they "name" their joint expectations. The participants make decisions, recognize jointly tasks, and assign duties. Small successes advance the process of decisions and acting, and create an identity in the group. This interaction base is strengthened with every new meeting.



Participatory Planning and GIS, Figure 6 Shared Knowledge Construction

Future Directions

The objective of participative planning processes is to develop a common goal in physical space. To develop a common goal it is necessary to get knowledge about social and physical usage of space. Knowledge construction, decision making, acting, and the bearing of responsibility are immediately connected within the planning steps of the planning group. The outcome of participative planning processes should be a special form of collaboration organizing social and physical usage of space. Therefore a shared knowledge construction is one important preconception for successful public participation [24].

Movement together through the space to be planned increases an *inter corporal* existence of the group. This inter corporal existence is the basic human experience of relationship, and contains all information of experiences and knowledge, and influences feeling, thinking, and acting patterns. In this inter corporal existence participants can feel, see, and interpret the actions and intentions of other participants.

Imagination and understanding emerges from the embodied experiences [25]. Human bodily movement and interaction integrate recurring patterns and develop new ones. It is possible to integrate information and transform it into knowledge in a mutual experiencing and understanding. Joint movement bring up joint experiences.

This concept is the backdrop for the assumptions about understanding, decision making, and knowledge construction processes among participating persons within a planning process. They integrate information and transform it into knowledge in a mutual understanding. Experiencing actions together leads to a shared knowledge construction. The knowledge the participants just brought with them, their constructions, their feeling and thinking patterns, remain. The concrete experience enlarges the shared knowledge [14]. The participants have to experience the concrete outcome of first decisions to strengthen the base of common action. Structures of social interaction are opened and renew the base of contact. The participants experience themselves mutually within every new meeting; they experience the concrete situation, make a common decision about this situation, and experience the first outcome of the decision.

Further the participants develop slightly changed roles and test them in new behaving patterns. This is the pre-condition: they are able to perceive new contents and information, and integrate it into their personal experience, knowledge, and acting. This flexible interaction enables to anticipate and imagine the future, to make decisions, to act, and to experience the outcome step by step.

How participants get more and more linked is shown in the next two figures.

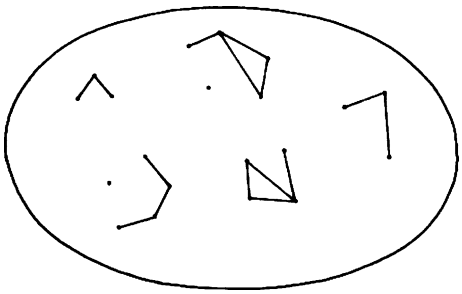
Figure 7 shows the usual structure of the whole group. The subgroups build by the participants relate to the social structure of the village.

After a meeting and after experiencing concrete situations participants are better linked [22].

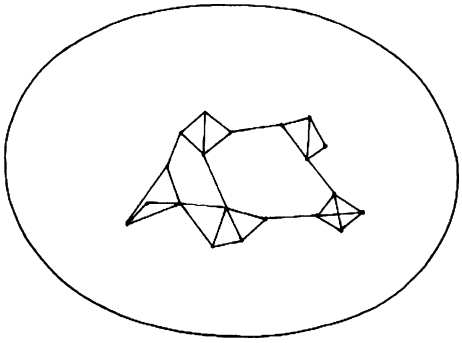
Key Applications

For planning issues GIS should support the context of *knowledge construction* of planners and participants to create meaning out of collaborative extraction of information, and to contribute to *group decisions* that use geospatial information.

To achieve effective group work with geospatial information it is essential to ask how the common experience of concrete situations can be combined with mediated decision making within a planning context. This suggests 1) that in virtual systems attention must be paid to the connection of the virtual situation to concrete previous experience of participants so that they are able to integrate new contents and information into their personal and shared experience, knowledge, acting, and experiencing of the outcome of acting, and 2) the interaction among par-



Participatory Planning and GIS, Figure 7 Initial Group Structure



Participatory Planning and GIS, Figure 8 Connection of subgroups

ticipants in virtual communities should be structured to achieve changing subgroups for building networks relating to changing tasks.

Cross References

► Decision-Making Effectiveness with GIS
► Geocollaboration

Recommended Reading

1. Argyle, M.: Bodyly Communication. Methuen & Co., London (1975)
2. Axelrod, R.: The Evolution of Cooperation. Basic Books, Princeton (1997)
3. Bakeman, R., Gottman, J.M.: Observing Interaction. Cambridge University Press, Cambridge, New York (1997)
4. Bales, R.F.: Interaction process analysis. Cambridge/Mass., Cambridge, New York (1951)
5. Bales, R.F.: Die Interaktionsanalyse. Ein Beobachtungsverfahren zur Untersuchung kleiner Gruppen. In: König, R. (ed.) Praktische Sozialforschung II. Beobachtung und Experiment in der Sozialforschung, pp. 148–167. Kiepenheuer & Witsch, Köln, Berlin (1962)
6. Buber, M.: Ich und Du. Reclam jun. GmbH & Co., Stuttgart (1995)
7. Damasio, A.: The Feeling of What Happens. First Harvest edition, San Diego, New York, London (1999)
8. Goffman, E.: Dramaturgie des Rollenhandelns. In: Retter, H.: Studienbuch Pädagogische Kommunikation. Klinkhardt, Bad Heilbrunn (2002)
9. Goldstein, K.: The Organism. Aholistic Approachto Biology Derived from Pathological Data in Man. Zone Books, New York (2000, 1934)
10. Hannaford, C.: Smart Moves: Why Learning is not all in your Head. Great Ocean Publishers, Arlington, VA (1995)
11. Haraway. D.: Situated Knowledges: The Science Question in Feminism and the Priviledge of Partial Perspective. In: Simians, Cyborgs, and Women. Routledge, New York, (1991)
12. Hochschild, J.: Where You Stand Depends on What You See: Connections among Values, Perceptions of Fact, and Prescriptions. In: Kuklinski, J. (ed.) Citizens and Politics: Perspectives from Political Psychology. Cambridge University Press, Cambridge (2000)
13. Jacobs, J.: The Death and Life of Great American Cities. Vintage Books, New York (1961)
14. Johnson, M.: The Body in the Mind. Chicago Press, Chicago (1987)
15. Lewin, K.: Field Theory in Social Science. Selected Theoretical Papers. D. Cartwright. Harper&Row, New York (1951)
16. Mansbridge, J.: Beyond Adversary Democracy. Basic Books, New York (1980)
17. Maturana, H.R.: Biologie der Realität. Suhrkamp, Frankfurt am Main (1998)
18. Merleau-Ponty, M.: Phänomenologie der Wahrnehmung. Walter De Gruyter & Co., Berlin (1966)
19. Pallaro, P. (ed.): Authentic Movement. Jessica Kingsley Publishers, London (1999)
20. Rottenbacher, C.: Presence in the Planning Process. GEOS 2004. Brasilien, (2004)
21. Rottenbacher, C.: Motion Increeases Emotional Correspondence in Geocollaboration. GIScience 2004. Baltimore, (2004)
22. Rottenbacher, C.: Emotional Correspondence Links. CORP 2005 (2005)
23. Rottenbacher, C.: Shared Knowledge Construction in Heterogeneous Groups. CORP 2006 (2006)
24. Smith, M.K.: Kurt Lewin, Groups, Experiental Learning and Action Research. The Encyclopedia of Informal Education. http://www.infed.org/thinkers/et-lewin.html (2001)
25. Varela, F.J., Thompson, E., Rosch, E.: The Embodied Mind. MIT Press, Cambridge Massachusetts (1997)
26. Young, M.I.: Inclusion and Democracy. Oxford Political Theory. Oxford University Press, Oxford, New York (2000)

Partitioning

► Geodemographic Segmentation

Path, Space-Time

► Time Geography

Pattern, Encounter

► Movement Patterns in Spatio-temporal Data

Pattern, Flock

► Movement Patterns in Spatio-temporal Data

Pattern, Leadership

► Movement Patterns in Spatio-temporal Data

Pattern, Moving Cluster

► Movement Patterns in Spatio-temporal Data

Pattern, Periodic

► Movement Patterns in Spatio-temporal Data

Pattern Recognition in Spatial Data

► Geographic Knowledge Discovery

Patterns

► Data Analysis, Spatial

Patterns, Complex

SANJAY CHAWLA
School of Information Technologies, The University of Sydney, Sydney, NSW, Australia

Synonyms

Correlated; Negatively-correlated; Spatial association; Co-location; Frequent itemset mining

Definition

Complex spatial relationships capture self, positive, negative and mixed relationships between spatial entities. These relationships abstract the commonly occurring relationships in ecology, cosmology and other disciplines where spatial referencing plays an important role. Table 1 enumerates the different types of complex relationships using the example of elliptical and spiral galaxy types from the field of cosmology. More formally, the definition of complex relationships is predicated on the concept of co-location.

Definitions 1 *(Co-location) Two spatial objects are said to co-locate if the Euclidean distance between the objects is less than or equal to the user-specified neighborhood distance d.*

Definitions 2 *(Positive) A positive relationship in spatial data is a set of features that co-locate at a ratio greater than some predefined threshold. In spatial data, the confidence of a positive relationship $A \rightarrow B$ is given by the fraction of unique Bs that co-occur in a clique containing the feature A.*

Definitions 3 *(Negative) A negative relationship in spatial data is defined as where a feature is absent from a given co-location at a ratio greater than a predefined threshold. Negative relationships are often denoted by "−".*

Definitions 4 *(Self-Co-location) A feature is defined as self-co-locating in spatial data if the items representing that feature co-locate with each other at a ratio greater than some predefined threshold. A Self-Co-location is denoted by "+".*

Definitions 5 *(Self-Exclusion) A feature is defined as self-excluding in spatial data if the items representing that feature co-locate with each other at a ratio less than some predefined threshold.*

Definitions 6 *(Complex) A complex relationship in spatial data is any relationship containing the properties of two or more of the other relationships.*

Historical Background

The study of spatial point processes is a core topic of research in the spatial statistics community [2,3,6]. The breakthrough in the data mining community is due to a series of papers by Shekhar et al. [8] and Huang et al. [4]. The two papers introduced a series of new *measures* which made it possible to efficiently discover colocation relationships in large spatial data sets. Furthermore, the *measures* introduced were closely related to the cross-K function [2]. This made it possible to *mine* for spatial relationships as opposed to *test* for them, which is the usual practice in Statistics.

Scientific Fundamentals

The methodology of discovering complex spatial relationships is based on a foundational data mining framework known as *frequent itemset mining*. The basic idea is as follows. Suppose there are n binary variables (also known as items or features) and the objective is to discover which elements of the power set of the n variables are correlated. A brute-force approach is computationally infeasible because the size of the search space is exponential in the number of variables. Instead, the elements of the power set can be examined in a leveled fashion and pruned as the power set lattice is being examined. The basic insight (known as the Apriori or anti-monotonic Property) is that

Patterns, Complex, Table 1 Types of Complex Spatial Relationships [5]

Relationship	Notation	Description	Example
Positive	$A \rightarrow B$	Presence of B in the neighborhood of A	Sa type Spiral Galaxies → Sb type Spiral Galaxies
Negative	$A \rightarrow -B$	Absence of B in the neighborhood of A	Elliptic galaxies tend to exclude spiral galaxies. E → −S
Self-Co-location	$A \rightarrow A+$	Presence of many instances of the same feature in a given neighborhood	Elliptic galaxies tend to cluster more strongly. E → E+
Self-Exclusion	$A \rightarrow -A+$	Absence of many instances of the same feature in a given neighborhood	Two or more of the same type of spiral galaxies are rarely found in the same neighborhood. Sa → −Sa+
Complex	$A+ \rightarrow -C,B$	combination of two or more of the above relationships	Clusters of elliptic galaxies tend to exclude other types of galaxies. E+ → −S

If a set of variables is not interesting, then neither are its supersets. This observation can be used to prune the power set lattice of variables in a computationally efficient manner.

The notion of *interestingness* is crucial in the frequent itemset mining framework. For example, the traditional measure of correlation does not satisfy the Apriori Property, but the simple COUNT function (known as the support) does. Huang et al. [4] have introduced the Maximal Participation Index (maxPI), which possesses a weak form of the Apriori Property and can be used to discover rare but interesting spatial relationships. Later, Arunasalam et al. [1] have formally and empirically shown how maxPI can be used to mine for complex spatial relationships. We briefly elaborate on the maxPI measure.

(Participation ratio) Given a co-location pattern L and a feature $f \in L$, the participation ratio of f, pr(L, f), can be defined as the support of L divided by the support of f. For example, in Fig. 1, the support of $\{A, B, C\}$ is 2 and the support of C is 6, so pr($\{A,B,C\},C$) = 2/6.

(Maximal Participation Index) Given a co-location pattern L, the maximal participation index of L, maxPI(L) can be defined as the maximal participation ratio of all the features in L, i. e., $\text{maxPI}(L) = \max_{f \in L}\{\text{pr}(L,f)\}$. For example, in Fig. 1,

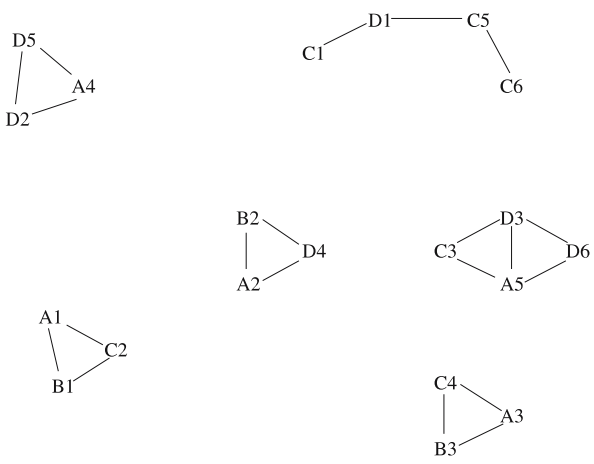
$$\text{maxPI}(\{A, B, C\}) = \max\left(\frac{2}{5}, \frac{2}{3}, \frac{2}{6}\right) = \frac{2}{3}.$$

A high maximal participation index indicates that at least one spatial feature (which we call the *maxfeature*) strongly implies the pattern. By using maxPI, rules with low frequency but high confidence can be found, which would otherwise be pruned by a support threshold.

Maximal participation index is not anti-monotonic with respect to the pattern containment relations. For example, in Fig. 1, $\text{maxPI}(\{A,C\}) = 3/5 < \text{maxPI}(\{A,B,C\}) = 2/3$. Interestingly, the maximal participation index does have the following weak monotonic property:

If P is a k-co-location pattern, then there exists at most one $(k-1)$ subpatterns P' of P such that $\text{maxPI}(P') < \text{maxPI}(P)$.

Complex relationships are not restricted to mining complex rules. Complex relationships can be used to provide stronger definitions and more accurate significance testing for simple relationships. In terms of confidence, the significance of a rule is given by the extent to which the observed confidence of a rule differs from the expected confidence given by a random distribution. Given a set of confident rules, the significance of these rules will depend on the relative size of the space from which they were taken. For example, Munro et al. [5] have shown that



Patterns, Complex, Figure 1 An Example of spatial co-location patterns

No	Clique
i	C_1, D_1
ii	C_5, D_1
iii	C_5, C_6
iv	A_4, D_2, D_5
v	A_5, C_3, D_3
vi	A_5, D_3, D_6
vii	A_1, B_1, C_2
viii	A_2, B_2, D_4
ix	A_3, B_3, C_4

Patterns, Complex, Table 2 Cliques in Fig. 1

The significance of a confident rule of the form $A \rightarrow B$ is independent of the self-co-location/exclusion of A, but is dependent on the self-co-location/exclusion of B.

Confidence is a measure of conditional probability. The confidence of $A \rightarrow B$ is $P(B|A)$. Thus, the probability of finding instances of B in a clique where A already exists is measured by the propensity of B'*s* to appear in the clique. This is the same if instances of A appear in one (self-colocation) or every clique (exclusion).

Key Applications

Traditional statistical techniques are designed for hypothesis testing: *Is the spatial relationship between two features significant?* In data mining, where large data sets with a multitude of features are the norm, the question of hypothesis generation is perhaps also interesting: *Find all multiple combinations of features such that the spatial relationship between these features is potentially significant*. Accurate and computationally efficient methods of discovering spatial relationships will help domain experts in diverse domains such as anthropology, cosmology, ecology, epidemiology, and geophysics. Additionally, many other disciplines unlock and discover new relationships and candidate theories in their respective fields.

Future Directions

One of the key computational challenges is to scale spatial collocation algorithms to handle an increasing number of spatial features. In traditional frequent mining, single features (items) can be pruned by themselves. However, because of the need to capture spatial relationships, single features cannot be pruned on their own. Thus, all pairs of features have to be initially computed leading to at least quadratic complexity in the number of features.

Another challenge is to extend the discovery the approach of mining for complex relationships to a spatio-temporal setting. As noted in Schabenberger et al. [7], observed point patterns are a snapshot of evolving patterns. For example, naturally generating oak trees initially tend to be clustered, then seem to be randomly distributed and finally tend to be arranged in a regular pattern as they compete for more space. Designing data mining techniques which can capture evolving trends in a spatio-temporal setting provides an exciting opportunity for future research.

Recommended Reading

1. Arunasalam, B., Chawla, S., Sun, P.: Striking two birds with one stone: Simultaneous mining of positive and negative spatial patterns. In: Proceedings of 2005 SIAM International Conference on Data Mining SDM05, to appear (2005)
2. Noel, A., Cressie, C.: Statistics for spatial Data. John Wiley and Sons, New York (1993)
3. Diggle, P.J.: Statistical Analysis of Point Processes. Chapman and Hall, New York (1983)
4. Huang, Y., Xiong, H., Shekhar, S., Pei, J.: Mining confident co-location rules without a support threshold. In: Proceedings of the 18th ACM Symposium on Applied Computing ACM SAC (2003)
5. Munro, R., Chawla, S., Sun, P.: Complex spatial relationships. In: Proceedings of the 3rd IEEE International Conference on Data Mining, ICDM 2003, pp. 227–234. IEEE Computer Society (2003)
6. Ripley, B.D.: Spatial Statistics. John Wiley and Sons, New York (1981)
7. Schabenberger, O., Gotway, C.: Statistical Methods for Spatial Data Analysis. Chapman and Hall, (2005)
8. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: a summary of results. In: Proceedings of the 7th International Symposium on Spatial and Temporal Databases SSTD01 (2001)

Patterns in Spatio-temporal Data

HUI YANG[1], SRINIVASAN PARTHASARATHY[2]

[1] Department of Computer Science and Engineering, San Francisco State University, San Francisco, CA, USA

[2] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA

Synonyms

Evolving spatial patterns; Spatio-temporal association patterns; Spatio-temporal object association

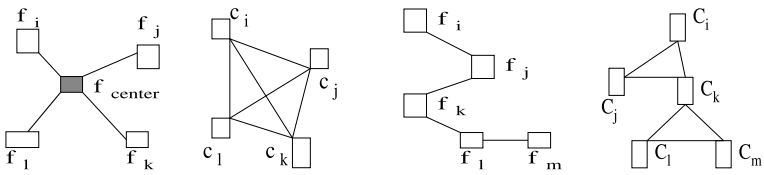
P

Definition

Spatio-temporal data refer to data that are both spatial and time-varying in nature, for instance, the data concerning traffic flows on a highway during rush hours. Spatio-temporal data are also being abundantly produced in many scientific domains. Examples include the datasets in computational fluid dynamics that describe the evolutionary behavior of vortices in fluid flows, and the datasets in bioinformatics that study the folding pathways of proteins from an initially string-like 3D structure to their respective native 3D structure.

One important issue in analyzing spatio-temporal data is to characterize the spatial relationship among spatial entities and, more importantly, to define how such a relationship evolves or changes over time. In the traffic flow example, one might be interested in identifying and monitoring the automobiles that are following one another far too close. Such an issue is often summarized as finding interesting spatio-temporal patterns.

A spatio-temporal pattern characterizes the spatial relationship among a collection of spatial entities and the evo-



Patterns in Spatio-temporal Data, Figure 1
Examples of spatial association patterns. **a** Star. **b** Clique. **c** Sequence. **d** minLink=2

lutionary behavior of such a relationship over time. As an example, Fig. 1 illustrates four types of spatial patterns, corresponding to four different types of spatial association. In this figure, each rectangle represents a spatial entity, and an edge indicates that the two involved entities hold a certain spatial relationship. For instance, an edge can mean that the Euclidean distance of the two entities is within a specified threshold. It can also mean that one entity is located to the left of the other. Or it can mean both of the above relations hold between the two entities. Assume that a collection of spatial entities $E = (e_0, e_1, \ldots, e_k)$ formed a star-like pattern (Fig. 1a) at time t_1 and continued in this fashion until time t_2. One can employ a spatio-temporal pattern in the form of $(Star, E, t_1, t_2)$ to effectively model such an evolving process.

Due to the following reasons, spatio-temporal patterns are often multifaceted. Furthermore, the spatio-temporal characteristics captured by such patterns often vary from application to application [6,9,10]:

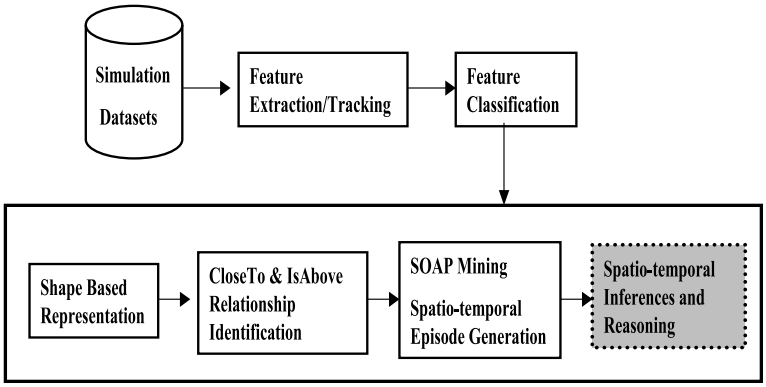
- Diversity of spatial relationship. For any pair of spatial entities, there exist a variety of spatial relations between them, such as directional relation, distance-based relation, and topological relation. Which of these relations should be captured in a spatio-temporal pattern is often specific to individual applications.
- Complexity of temporal relationship. For instance, there exist 13 possible relations between two time intervals [1]. Again, it is often governed by the applications to decide what relations should be considered in the spatio-temporal patterns.
- Representation of spatial entities: points or geometric objects?
- Varying application-specific requirements. For instance, one application might require one to capture how the distances between entities change in time, whereas another application might be interested in investigating both the distance and relative directional arrangement between entities.

Note that evolving spatial clusters—collections of spatial entities that are similar to one another (e. g., entities within the same vicinity)—are another type of spatio-temporal patterns. The main difference between evolving spatial clusters and the above-described spatio-temporal patterns resides in the number of involved spatial entities. A spatial cluster often consists of much more spatial entities than a spatio-temporal pattern. Additionally, spatio-temporal patterns are more versatile in the sense that a variety of spatial and temporal relations can be considered simultaneously as needed, whereas spatial clusters are often concerned about only the distance-based relationship among entities.

Historical Background

The history of spatio-temporal association patterns is closely related to that of spatial association patterns, since the former is often derived by incorporating the temporal dimension into the latter. Spatial association patterns were first studied by Koperski and Han [4]. This early work focuses on extracting patterns specified in advance. Following this work, a considerable amount of work was conducted to detect spatial clusters [2]. Such clusters mainly captured the spatial proximity among entities. Driven by the widespread location-based services at the turn of the century, researchers started to take a special interest in identifying spatial patterns that involve a smaller set of entities within a confined spatial neighborhood [6]. Such patterns were later termed as spatial collocation patterns [3]. For instance, the collocation pattern *(weather, airline schedule, Starbucks coffee shops)* captures the phenomenon that the customers at Starbucks coffee shops tend to request weather information and airline schedules together through cellular phone. However, the research work up to this point often simplified spatial entities to point objects and mainly considered the Euclidean distance between objects. Recently, several studies were carried out to overcome such limitations [9]. In these studies, spatial entities are represented as geometric objects of different shape and size. In addition, the spatio-temporal patterns are capable to capture multiple spatial relations (e. g., both distance-based and directional relation). Consequently, the term spatial or spatio-temporal object association patterns were coined to emphasize such facts [10].

Another prominent development of spatio-temporal patterns analysis is that it has found more and more applications in scientific domains, such as astronomy, meteorology, biochemistry, and bioinformatics. This is in contrast to its earlier application mainly in geographic information systems.



Patterns in Spatio-temporal Data, Figure 2
A Generalized Framework for Analyzing Spatio-temporal Scientific Data

Scientific Fundamentals

The process of identifying spatio-temporal patterns can be decomposed into three main phases. The first phase is data preprocessing. Main tasks in this phase include the following: (1) Determine the representation scheme of spatial entities: points or geometric objects? If it is the latter case, what geometric properties and domain specific attributes need to be considered? (2) Concretize the spatio-temporal patterns: what spatial and temporal relations should the patterns be modeling? (3) Identify and define the measurements that measure the "interestingness" of a pattern. For instance, *support* and *prevalence* have been proposed by Yang *et al.* to characterize the significance of a pattern [10]. The second phase is to efficiently and effectively discover interesting spatio-temporal patterns. One main challenge is to achieve good scalability and performance in the presence of a large volume of data, which are often in the range of gigabytes and even terabytes. Efficient data structures and optimization strategies are often employed towards improving scalability and performance. The third and final phase is to evaluate the identified spatio-temporal patterns and put them into use. The nature and implementation of this phase is often application-specific.

In scientific domains, the discovery of spatio-temporal patterns often brings up new challenges. For instance, to discover spatio-temporal patterns of vortices in fluid flows, one needs to first detect and extract vortical objects at different time. This task by itself is still under intensive study currently. Readers are referred to [10] for more details on a generalized framework for analyzing scientific spatio-temporal data. This framework is illustrated in Fig. 2.

Key Applications

Spatio-temporal association patterns have been used to address various issues in many domains. Below is a list of representative applications from different domains.

Traffic Management

Spatio-temporal association patterns can be used to identify and predict potential accidents by modeling automobiles within dangerous distance. Such patterns can also be used to redirect traffic flows, thereby avoiding potential traffic jam.

Behavior Tracking in Security Surveillance Systems

Surveillance systems track and record the behavior of human subjects aiming at identifying suspicious behaviors. One can use spatio-temporal patterns to model such behaviors by associating a person's movement with objects in the surrounding area.

Astronomy

In astronomy, spatio-temporal patterns can be used to capture the evolution of interactions among astronomical objects in the vicinity by exploring the data accumulated in the past.

Transmissible Disease Control

To control and predict the spreading rate of transmissible diseases (e. g., SARS), one critical issue is to have a clear notion of how people in the infected areas regularly relate to each other and with people in the disease-free areas. Spatio-temporal association patterns can be applied to model such people-people interactions.

Computational Molecular Dynamics: Interaction and Evolution of Defects in Materials

It has been observed that multiple defects in materials often interact with each other. Such interactions eventually might lead to undesirable results, such as the amalgamation of small defects and the breakdown of large defects. Again, such behavior can be modeled and captured by identifying spatio-temporal association patterns of defects.

Computation Fluid Dynamics: Characterizing Vortical Flows

Vortices—swirling regions around a common center—in vortical flows can often produce undesirable effects, especially when such vortices interact with one another. For instance, vortices in the air flows surrounding an airplane can lead to audible noise and strong vibration. Therefore, designers often resort to computer simulations to study vortical flows around a certain model. Here one can use spatio-temporal patterns to characterize the evolving behavior of vortices at different locations of the model under study.

Bioinformatics: Protein Folding Trajectories Analysis

A protein folding trajectory describes the folding path of a protein from an initially string-like structure to its final native and often complex structure. Along this path, amino acids, the building blocks of a protein, interact with one another. Such interactions often result in a variety of folding events, such as nucleation and secondary structure formation. It has been demonstrated that spatio-temporal association patterns could be applied to address several issues: (1) summarizing a folding trajectory; (2) detecting and ordering folding events along a trajectory; and (3) identifying a consensus partial folding pathway across different trajectories of a protein [11].

Future Directions

Discovering interesting and meaningful spatio-temporal association patterns is still a relatively new problem. Below are several potential research focuses related to this problem: (1) design scalable algorithms that can handle large volume of spatio-temporal datasets. Candidate solutions include the following: integrating efficient indexing schemes in the process and developing parallel or distributed algorithms; (2) implement effective approaches to incorporate domain-specific knowledge in the pattern discovering process; (3) utilize visualization techniques to facilitate an easier verification and a better understanding of the discovered spatio-temporal patterns; and (4) implement generalized software systems to discover spatio-temporal patterns similar application domains.

Recommended Reading

1. Allen, J.F.: Maintaining knowledge about temporal intervals. Commun. ACM. **26**(11), 832–843 (1983)
2. Ester, M., Kriegel, H.P., Sander, J.: Algorithms and applications for spatial data mining. Geographic Data Mining and Knowledge Discovery, Research Monographs. In: GIS Chapter 7 (2001)
3. Huang, Y., Xiong, H., Shekhar, S., Pei, J.: Mining confident co-location rules without a support threshold. In: SAC 03: Proceedings of the 2003 ACM symposium on applied computing, pp. 497–501. ACM Press (2003)
4. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In SSD 95: Proceedings of the 4th International Symposium on Advances in Spatial Databases, pp. 47–66. Springer-Verlag (1995)
5. Mokbel, M.F., Ghanem, T.M., Aref, W.G.: Spatio-temporal access methods. Technical report, Department of Computer Sciences, Purdue University
6. Morimoto, Y.: Mining frequent neighboring class sets in spatial databases. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 353–358. ACM Press (2001)
7. Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K.: Detection of emerging space-time clusters. In: Proceedings of SIGKDD 2005, pp. 218–227 (2005)
8. Rao, C.R., Suryawanshi, S.: Statistical analysis of shape of objects based on landmark data. Proc Natl Acad Sci U S A. **93**(22), 12132–12136 (1996)
9. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., Yoo, J.S.: A framework for discovering co-location patterns in data sets with extended spatial objects. SIAM Intl. Conf. on Data Mining (SDM), April 2004
10. Yang, H., Parthasarathy, S., Mehta, S.: A generalized framework for mining spatio-temporal patterns in scientific data. In KDD 2005: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 716–721. ACM Press, New York, NY, USA (2005)
11. Yang, H., Parthasarathy, S., Ucar, D.: A spatio-temporal mining approach towards summarizing and analyzing protein folding trajectories. Algorithms Mol. Biol. **2**(3) (2007)

Peano Curve

▶ Indexing of Moving Objects, B^x-Tree

Peer Data Management

▶ Database Schema Integration

Peer to Peer

▶ Cloaking Algorithms for Location Privacy

Peer-to-Peer Caching for Spatial On-Line Analytical Processing

▶ Olap Results, Distributed Caching

Peer-Tree (Spatial Index)

▶ Data Collection, Reliable Real-Time

Perceptory Pictograms

► Modeling with Pictogrammic Languages

Personalization

► Geospatial Semantic Web: Personalisation
► User Interfaces and Adaptive Maps

Personalized Maps

► Mobile Usage and Adaptive Visualization

Personalized Visualization

► Mobile Usage and Adaptive Visualization

Phantom Update Protection

► Concurrency Control for Spatial Access Method

Phenomenon Spatial Field

► Geosensor Networks, Estimating Continuous Phenonena

Photogrammetric Applications

THOMAS LUHMANN
Institute for Applied Photogrammetry
and Geoinformatics, Oldenburg, Germany

Synonyms

Photogrammetry; Application; Aerial; Close range; Data acquisition; 3D city models

Definition

The numerous application areas for photogrammetry can be categorized into two major groups: topographic applications based on aerial and satellite imagery, and close-range applications.

Historical Background

Photogrammetry was invented between 1850 and 1860 independently by the French Laussedat and the German Meydenbauer. Aerial photogrammetry became a focus with the development of aircraft and fast film material. It was further developed by the coming of color film and, in the second half of the twentieth century, by satellite imaging systems. The photogrammetric industry developed analog optical and mechanical stereo plotting devices and image rectifiers to meet the practical demands. Since the 1960s, computer-based analytical processing methods have been used intensively. Consequently, digital photogrammetry became state of the art in the 1990s, when digital imagery and digital imaging devices came into use. In close-range photogrammetry, the main fields of application were addressing architecture and cultural heritage, accompanied by a variety of very specialist applications and solutions. With the development of self-calibrating bundle adjustment programs around 1980, industrial photogrammetry was strongly used mainly for large scale metrology, e. g., for antennae or in the aerospace industry. Again, the development of high-resolution digital cameras pushed the technique further ahead. Nowadays, photogrammetry is an accepted and widely used tool in many industrial, medical, and engineering tasks.

Scientific Fundamentals

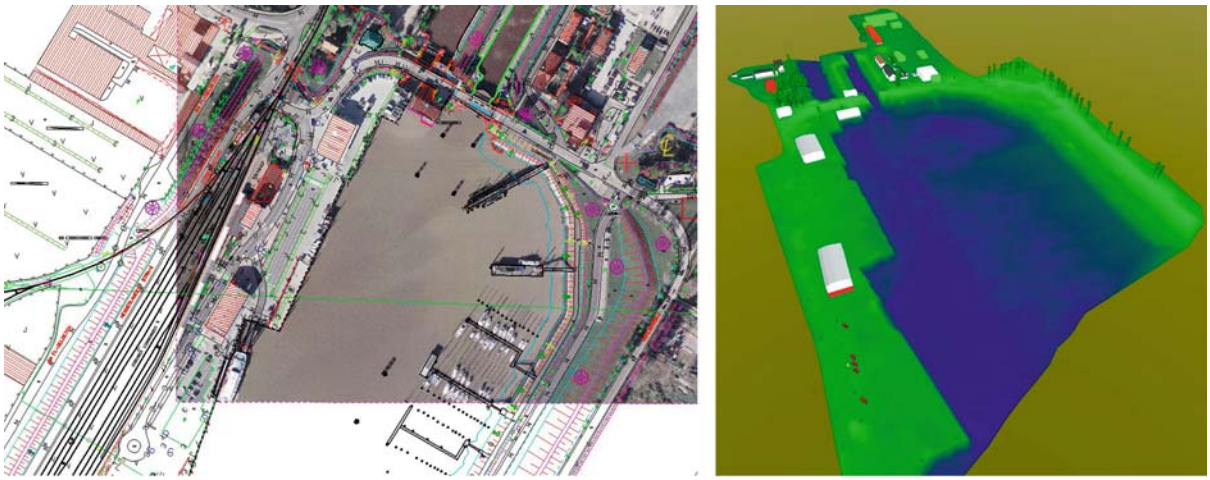
Please refer to the entries on Mathematical Concepts of Photogrammetry, Photogrammetric Products, and Photogrammetric Sensors.

Key Applications

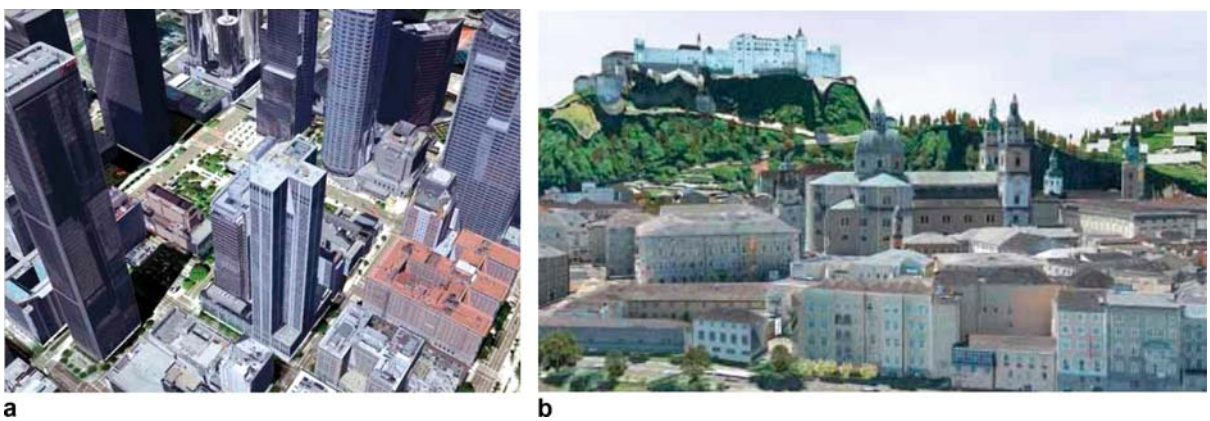
Aerial Applications

Applications in aerial photogrammetry can be characterized by often similar imaging configurations, i. e., equal or similar cameras, nadir imagery, large imaging distances (flying height) and image scales usually between 1:2,000 and 1:30,000. The most important products of aerial photogrammetry are orthophotos, 3-D terrain and city models, and vector data usually used as input for geographic information systems (GIS).
The following examples show a small spectrum of the applications in aerial photogrammetry.

GIS Data Acquisition Photogrammetric data generation for GIS purposes is the most important application of aerial photogrammetry. An example of photogrammetric data acquisition and modeling for a GIS application is shown in Fig. 1. The task was the extraction of data about the harbor of the German town Emden in order to provide 2-D and 3-D information for an internet-based information system. Besides aerial photographs, additional data sources such as cadastral maps, sonar depth measurements and terrestri-



Photogrammetric Applications, Figure 1 Photogrammetric data extraction and modeling for a harbor information system. **a** Superimposition of **geographic information systems** (GIS) data and aerial image. **b** 3-D depth model of port basin

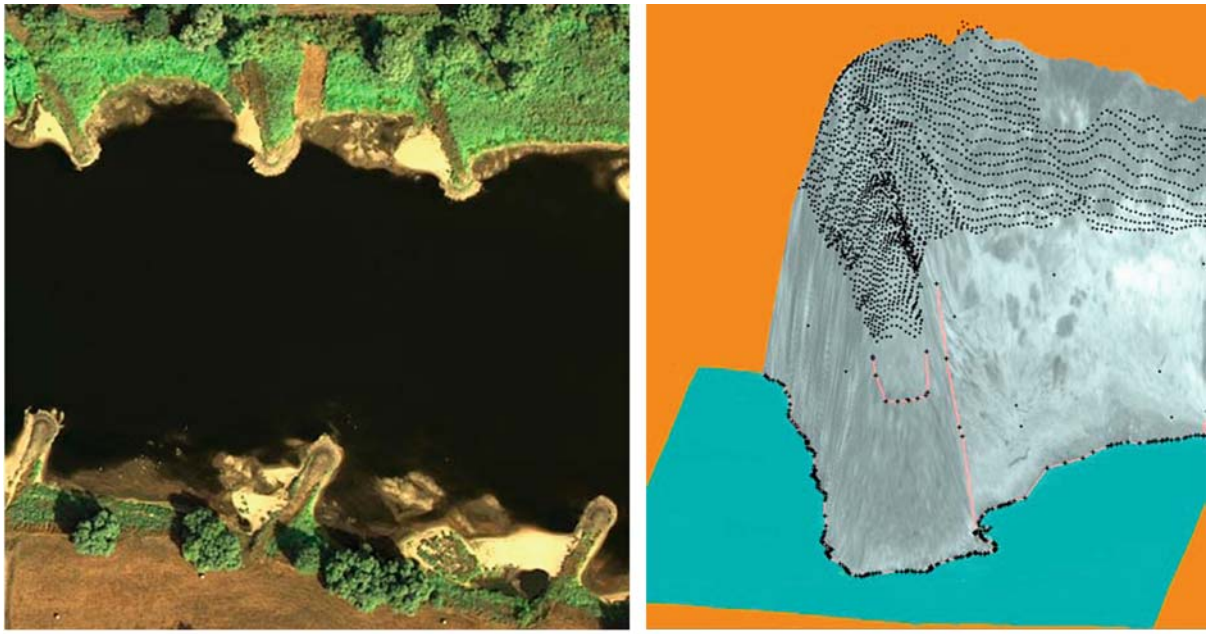


Photogrammetric Applications, Figure 2 Examples of 3-D city models from Los Angeles and Salzburg (CyberCity)

al images have been used. Photogrammetric data compilation is performed using a digital stereo workstation.

3-D City Models Aerial images are the most important data source for measuring 3-D city models. Image information is used to extract 3-D points and topological information (wire frame models). In addition, the operator can identify different types of buildings and other objects in order to classify the scene in terms of database attributes. 3-D city models are increasingly used for touristic purposes, urban planning, real estate management and emission monitoring and prediction. Figure 2 shows examples of city models that have been extracted from aerial imagery interactively. The facades of the buildings can be textured from terrestrial images. For visualization of the resulting huge amount of data, specialized visualization software and data structures are employed.

Flood Monitoring Monitoring of rivers is becoming increasingly important for environmental protection and flood disaster management. Aerial or high-resolution satellite imagery give fast access to regional terrain and flood information, often in combination with additional GIS data such as terrain models and water resources. As an example, heavy floods have repeatedly affected the German river Elbe region. Precise and up-to-date maps are therefore indispensable for the prediction of water levels and streams, but also for disaster management and rebuilding. Figure 3 shows an aerial image of the Elbe river from August 2003 taken by a digital aerial camera ZI DMC.



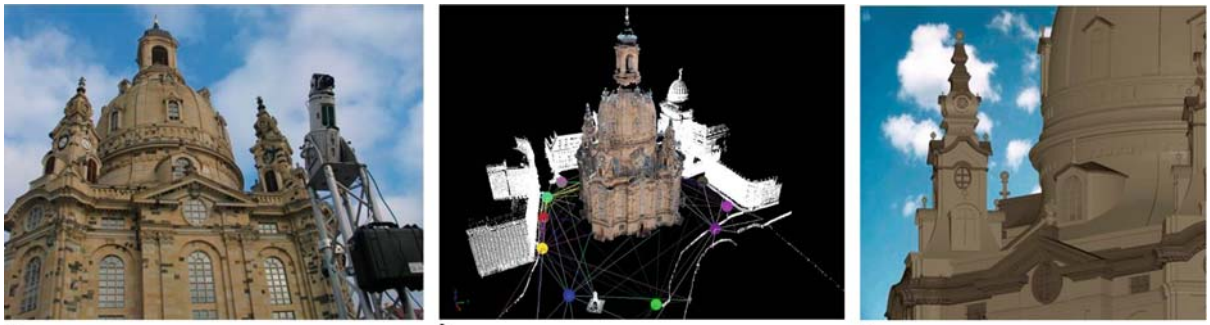
Photogrammetric Applications, Figure 3 Original image (**a**) and reconstructed groyne (**b**) (BfG Koblenz, EFTAS Münster)

Photogrammetric Applications, Figure 4 Original images (**a–i**) and resulting orthophoto mosaic (**j**)(IAPG Oldenburg)

These images have been used to measure digital terrain models of the river banks in order to complement airborne laser-scanning data which was acquired during high water. As an example, Fig. 3 right shows a 3-D model of a groyne as part of a larger terrain model that has been reconstructed from imagery.

Close-Range Applications

In contrast to aerial photogrammetry, the application areas in close-range photogrammetry are much broader. The most common applications address architecture and cultural heritage, industrial production control and quality assur-

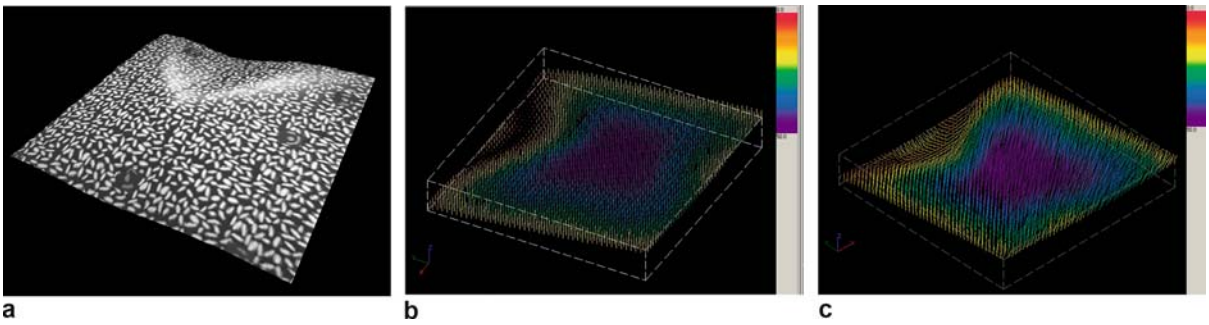


Photogrammetric Applications, Figure 5 Original image (**a**), 3-D point cloud (**b**) and example visualization (**c**) (Riegl)

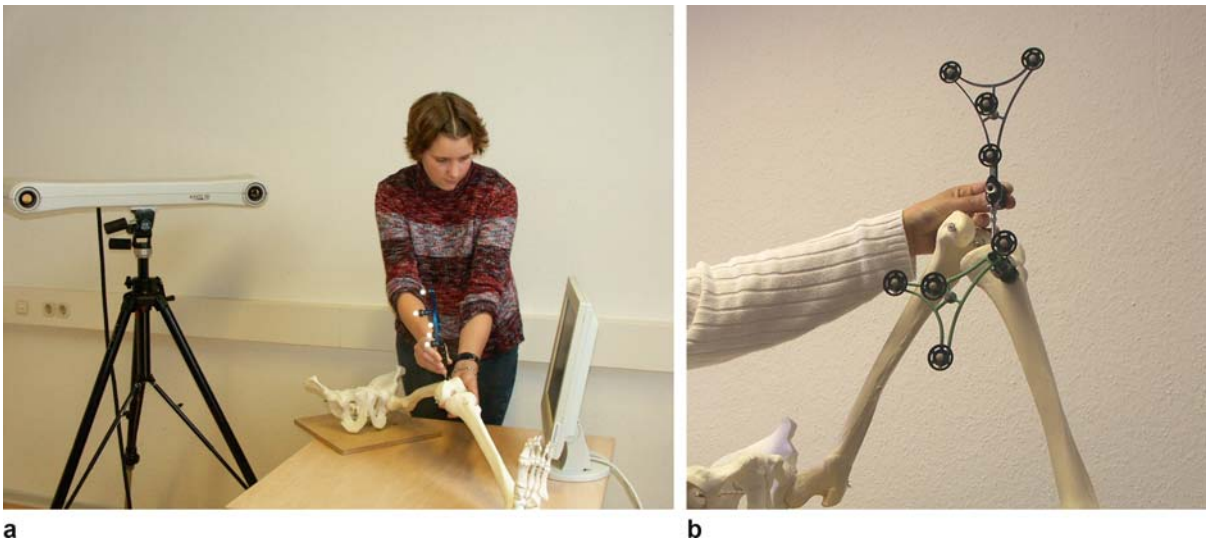
Photogrammetric Applications, Figure 6 Accident scene recording and photogrammetric processing (Photometrix)

ance, medicine, forensic and scientific applications. Imaging systems range from simple video to high-resolution digital cameras, from panoramic to high-speed cameras. Cameras are configured as single cameras or as multicamera setups. Imaging systems for the close range are often not designed as highly stable metric cameras, and hence they often must be calibrated either at shorter time intervals, or simultaneously with bundle adjustment for object reconstruction. The hybrid combination of cameras with other sensors, e. g., terrestrial laser scanners, is of increasing interest.

Close-range systems can be categorized into online and offline systems. Online systems generate 3-D data within a continuous data flow directly on the object site, e. g., for



Photogrammetric Applications, Figure 7 **a–c** High-speed image sequence processing for dynamic deformation analysis



Photogrammetric Applications, Figure 8 Stereo navigation system (**a**) and hand-held probe (**b**)

the navigation of tools in medical applications, or for the control of machines and robots in industry. In offline systems, image acquisition is often separated from image processing and object reconstruction. Hence, both parts can be performed in different locations, at different times and by different people. Products of close-range photogrammetry range from simple 3-D coordinates to process parameters, from free-form surface models to animated 3-D objects in static or dynamic environments.

The following examples cover four major application areas of close range photogrammetry, namely architecture, forensic analysis, industry, and medicine. Additional examples are given in [1,2,3,4,5,6,7].

Visualization of Architectural Objects Recording and visualization of buildings, archaeological sites or cultural heritage objects is one of the traditional photogrammetric applications. Besides 2-D drawings and plans, an increasing demand on rectified orthoimagery and 3-D models can be observed. Figure 4 shows a high-resolution image mosaic that consists of nine digital images with $4{,}000 \times 4{,}000$ pixels each.

Through a combination of close-range photogrammetry and terrestrial 3-D laser scanning it is possible to measure the complex surface shapes that exist in diverse forms for buildings, industrial process plants, archaeological excavations and sculptures. The example in Fig. 5 shows the Frauenkirche in Dresden, recorded by digital images, as well as airborne and terrestrial laser scans from multiple survey stations. The end result is a realistic, textured 3-D model of the church created by 3-D monoplotting.

Accident Recording Recording of traffic accidents is often characterized by difficult imaging configurations such as weak intersections of image rays or complex object scenes. The desired use of consumer digital cameras

requires powerful and robust image calibration and orientation. Figure 6 shows an example of an accident scene and a subsequent scene reconstruction processed by the iWitness (Photometrix) software package.

Dynamic Surface Reconstruction Figure 7 shows the results of a dynamic deformation analysis of a car body part. The scene has been recorded by two high-speed cameras with a frame rate of 1,000 Hz. The object surface has been prepared by an artificial pattern in order to provide sufficient image texture for matching. In each epoch the surface is reconstructed by stereocorrelation following physical surface points through the image sequence.

Medical 3-D Navigation Figure 8 displays a stereocamera system based on two video cameras (AXIOS 3D Services) that is used in medical applications for the measurement of the body and navigation of tools for computer assisted surgery. Usually this kind of system guides the surgeon for precise handling of surgical tools with respect to other tools, or part of the human body. The typical accuracy ranges from 3 mm down to 0.3 mm in a 1 m^3 measurement volume.

Future Directions

Photogrammetry serves as a flexible measurement tool in many different application fields. For both major areas, namely geotechnology and close-range applications, significant market growing rates of 15% per year and more have been predicted. It is therefore obvious that photogrammetry and 3-D image processing are fundamental upcoming technologies for a broad variety of applications.

Cross References

- ▶ Photogrammetric Products
- ▶ Photogrammetric Sensors
- ▶ Visualizing Constraint Data

Recommended Reading

1. Atkinson, K.B. (ed.): Close Range Photogrammetry And Machine Vision. Whittles, Caithness, UK (1996)
2. Fryer, J.G., Mitchell, H.L., Chandler, J.H.: Applications of 3D Measurement From Images. Whittles, Caithness, UK (2006)
3. Karara, H.M. (ed.): Non-Topographic Photogrammetry. 2nd edn. American Society for Photogrammetry and Remote Sensing. Falls Church, Virginia (1989)
4. Luhmann, T. (ed.): Nahbereichsphotogrammetrie in der Praxis. Wichmann, Heidelberg (2002)
5. Luhmann, T., Robson, S., Kyle, S., Harley, I.: Close-Range Photogrammetry. Whittles, Caithness, UK (2006)
6. Madani, M.: Photogrammetric applications. In: McGlone (ed.) Manual of Photogrammetry, 5th edn., pp. 1015–1104. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland (2004)>
7. ISPRS Commission V Close-Range Sensing: Analysis and Applications. http://www.commission5.isprs.org/. Accessed 20 Aug 2007

Photogrammetric Cameras

▶ Photogrammetric Sensors

Photogrammetric Images

▶ Photogrammetric Sensors

Photogrammetric Methods[1]

OLAF HELLWICH
Computer Vision and Remote Sensing,
Berlin University of Technology, Berlin, Germany

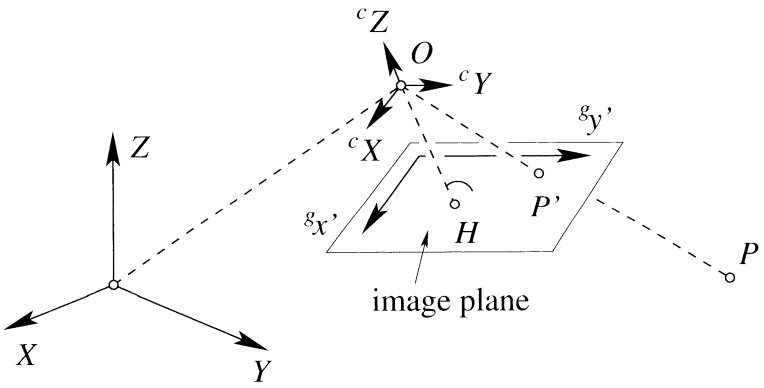
Synonyms

Camera model; Sensor orientation; Object reconstruction; Methods of photogrammetry; Single image; Image pair; Image triplet; Multiple-image bundle block; Bundle adjustment; Central projection; Central perspective; Fundamental matrix

Definition

The classical task of photogrammetry is the recovery of information from images of a scene [1]. This entry mathematically describes the geometry of single-perspective images, image pairs, image triplets, and blocks of several images. The notion "image" is very broad. Most widely used are images of frame cameras, i. e., the image is a 2D mapping from 3D object to 2D image space where the entire frame is exposed simultaneously through a lens. It is assumed that a unique projection center exists, so that the light rays between object and image points pass through a single point. In some cases, in particular when generating 2D images with a sweeping line camera, a unique projection center for the whole image does not exist. However, in all cases one assumes that one is able to determine a projection ray for a measurable image point in order to infer 3D information in object space from 2D measurements.

[1] This entry summarizes contents from the Manual of Photogrammetry [5]



Photogrammetric Methods, Figure 1 Mapping with a digital camera, object coordinate system $[X, Y, Z]$, projection center O, camera coordinate system $[{}^{c}X, {}^{c}Y, {}^{c}Z]$, image or sensor coordinate system $[{}^{g}x', {}^{g}y']$, principal point H, object point P, image point P' [5]

Historical Background

In 1492 Leonardo da Vinci graphically demonstrated optical projection. Albrecht Dürer constructed mechanical devices to do perspective drawings of natural and studio scenes. In his classical treatise *The Free Perspective*, Henry Lambert dealt with the concept of inverse central perspective and space resection of conjugate image rays. It contained the geometric fundamentals of the process that 100 years later was named photogrammetry. With prophetic insight Guido Schreiber had rendered a treatise in 1829 on *The Process and Formulae for Air Topographic Equations and Determination of the Camera Station*, envisioning the time when Earth's image would be produced from a bird's-eye view. In 1849 Aimé Laussedat, an officer in the Engineering Corps of the French Army, embarked upon a determined effort to prove that photography could be used with advantage in the preparation of topographic maps. His work in this field was so complete that the principles demonstrated by practical applications are still in use. Not much later Ernst Abbe, cofounder of the Zeiss Works, placed the design of optical elements and their combination on a rigorous mathematical basis. In 1893, Albrecht Meydenbauer published a paper on the new method of photographic surveying in which the first use of the word photogrammetry appears. By the end of the 1930s, the semicomputational processes of the early days had fully been replaced by the optomechanical process of orienting stereo imagery to form a stereo model. This situation should change again with the advent of computers. After World War II, Hellmut Schmid developed the principles of multistation analytical photogrammetry. He rigorously applied the least-squares method to the simultaneous orientation of any number of photographs with a complete study of error propagation. In the 1960s there was considerable activity in developing and implementing practical adjustment algorithms for aerotriangulation, e. g., Duane Brown came up with an elegant and general treatment of least-squares adjustment and error propagation leading to computer programs, e. g., for extraterrestrial missions like Apollo. In the 1990s concepts of algebraic projective geometry were used to derive general direct solutions for photogrammetric problems advantageous in the automation of image analysis using uncalibrated low-cost cameras.

Scientific Fundamentals

In the following, a geometric model of the projection of points into the image generated by a real camera is formulated. It allows the projection process to be inverted to infer the spatial direction to 3D points from their observed images, and to use this to determine the spatial position of the camera and the 3D position of the observed points [2,3].

For modeling the projection, points are represented in three coordinate systems (Fig. 1): the object coordinate system S_o with object coordinates $\boldsymbol{x} = (X, Y, Z)^t$, the camera coordinate system S_c with camera coordinates ${}^{c}\boldsymbol{x} = ({}^{c}X, {}^{c}Y, {}^{c}Z)^t$, and the sensor coordinate system S_g with image coordinates ${}^{g}\boldsymbol{x}' = ({}^{g}x', {}^{g}y')^t$. It is assumed that all coordinate systems are Euclidean and right handed.

The exterior orientation transforms the coordinates $\boldsymbol{x}_P$ of a point P from the object coordinate system S_o into the camera system S_c. This can be achieved in two steps by a translation of the object coordinate system S_o into the projection center O, and a rotation of the coordinate system S_o into the system S_c. The rotation matrix $\boldsymbol{R}$ can be represented by three independent parameters. In Euclidean coordinates:

$$ {}^{c}\boldsymbol{x}_P = \boldsymbol{R}(\boldsymbol{x}_P - \boldsymbol{x}_O) \,. \tag{1} $$

Often, camera models are formulated using homogeneous coordinates. Homogeneous coordinates $\mathbf{e}$ of an entity are invariant with respect to multiplication by a scalar $\lambda \neq 0$, thus that $\mathbf{e}$ and $\lambda\,\mathbf{e}$ represent the same entity. For instance, a 3D point with Euclidean coordinates $\boldsymbol{x} = (X, Y, Z)^t$ has

P

homogeneous coordinates $\mathbf{x} = (U,\ V,\ W,\ T)^{\mathrm{t}}$ which are related by:

$$\mathbf{x} = \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} = \begin{bmatrix} U \\ V \\ W \\ T \end{bmatrix} = \begin{bmatrix} XT \\ YT \\ ZT \\ T \end{bmatrix}.$$

In homogeneous coordinates (1) reads:

$$\begin{aligned} {}^{c}\mathbf{x}_P &= \begin{bmatrix} {}^{c}\boldsymbol{x}_P \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R} & \mathbf{0} \\ \mathbf{0}^t & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{x}_0 \\ \mathbf{0}^t & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_P \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{R} & -\boldsymbol{R}\boldsymbol{x}_0 \\ \mathbf{0}^t & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_P \\ 1 \end{bmatrix} = {}^{c}\mathbf{M}\mathbf{x}_P . \end{aligned} \tag{2}$$

When mapping with an ideal central perspective camera having a distortion-free lens and a planar sensor area, the Euclidean sensor coordinate system S_{g} is centered at the point on the image plane closest to the projection center, i. e., at the principal point, and the axes of this system are parallel to the axes of the camera coordinate system S_{c}. Then the homogeneous coordinates of an image point are:

$$\begin{aligned} {}^{c}\mathbf{x}'_{\overline{P}} &= \begin{bmatrix} {}^{c}u'_{\overline{P}} \\ {}^{c}v'_{\overline{P}} \\ {}^{c}t'_{\overline{P}} \end{bmatrix} \\ &= \begin{bmatrix} c & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} {}^{c}X_P \\ {}^{c}Y_P \\ {}^{c}Z_P \\ 1 \end{bmatrix} = {}^{c}\mathbf{P}_c\, {}^{c}\mathbf{x}_P . \end{aligned}$$

The 3×4 matrix ${}^{c}\mathbf{P}_{\mathrm{c}}$ performs the projection from the object point P given in the camera coordinate system into the point $\overline{P\prime}$ in an ideal sensor coordinate system. For the ideal camera ${}^{c}\mathbf{P}_{\mathrm{c}}$ contains only one parameter defining its interior orientation, namely its principal distance c.

Using (1) and (2), the composed mapping from object space to image space with an ideal camera is expressed as:

$$\begin{aligned} {}^{c}\mathbf{x}' &= {}^{c}\mathbf{P}\mathbf{x} = {}^{c}\mathbf{P}_c\, {}^{c}\mathbf{M}\mathbf{x} \\ &= \begin{bmatrix} c & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & -\boldsymbol{R}\boldsymbol{x}_O \\ \mathbf{0}^t & 1 \end{bmatrix} \mathbf{x} . \end{aligned}$$

Introducing a calibration matrix:

$${}^{c}\mathbf{K} = \begin{bmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

the projection reads

$${}^{c}\mathbf{x}' = {}^{c}\mathbf{K}\boldsymbol{R}\,[\mathbf{I}\,|{-}\boldsymbol{x}_O]\,\mathbf{x}$$

The Euclidean coordinates of the image point are given by the so-called collinearity equations:

$${}^{c}x' = c\frac{r_{11}(X-X_O) + r_{12}(Y-Y_O) + r_{13}(Z-Z_O)}{r_{31}(X-X_O) + r_{32}(Y-Y_O) + r_{33}(Z-Z_O)}$$

$${}^{c}y' = c\frac{r_{21}(X-X_O) + r_{22}(Y-Y_O) + r_{23}(Z-Z_O)}{r_{31}(X-X_O) + r_{32}(Y-Y_O) + r_{33}(Z-Z_O)}$$

derived by dividing through the third component.

In order to model a real camera the projection model is extended in two steps:

- All terms which guarantee the projection to be straight-line-preserving are added.
- Then additional terms allowing the modeling of general cameras still having a unique projection center are introduced [4].

Note that the ideal point $\overline{P'}$ is assumed to be identical to the measurable point P' observed in a skew coordinate system being related to the ideal coordinate system by an affine transformation. The parameters of this transformation are the translation of the coordinate system into the principal point $\left(x'_H, y'_H\right)^t$ of the sensor coordinate system S_{g}, the correction of the scale of the y' coordinates with respect to the x' coordinates by the factor $1+m$, and the shear of the ${}^{c}y'$ axis $s = \tan(\alpha)$, where α is the shear angle. Including this transformation into the calibration matrix results in:

$$\mathbf{K} = \mathbf{H}_c\, {}^{c}\mathbf{K} = \begin{bmatrix} c & cs & x'_H \\ 0 & c(1+m) & y'_H \\ 0 & 0 & 1 \end{bmatrix}.$$

The final projection then reads as:

$$\mathbf{x}' = \mathbf{P}\mathbf{x} \tag{3}$$

with the homogeneous projection matrix:

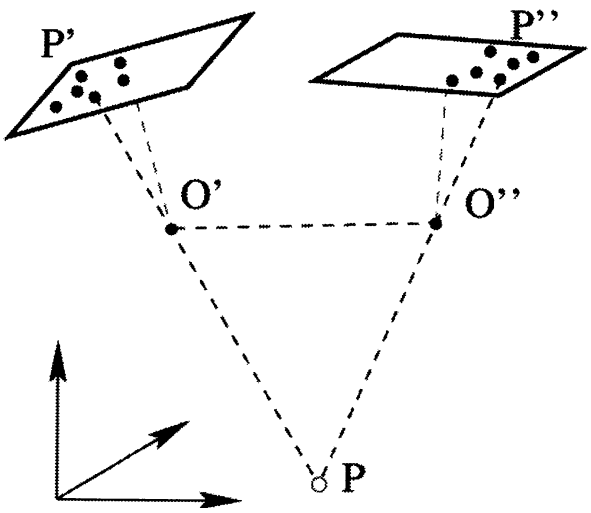
$$\mathbf{P} = \mathbf{K}\boldsymbol{R}\,[\mathbf{I}\,|{-}\boldsymbol{x}_O]$$

which contains 11 parameters, namely the 5 parameters of the interior orientation in matrix $\mathbf{K}$ and the 6 parameters of the exterior orientation.

The mapping Eq. 3 with the elements p_{ij} of P is explicitly given as:

$$\begin{aligned} x' &= \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}} \\ y' &= \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}} . \end{aligned}$$

It is called the direct linear transformation (DLT) [5], as it directly relates the Euclidean coordinates of the object points and measurable sensor coordinates of the image points of a straight-line-preserving camera.



Photogrammetric Methods, Figure 2 Geometry of an image pair [1]

In the above-mentioned second step an additional homography-like transformation depending on the local position ${}^{c}\boldsymbol{x}'$ in the image and additional parameters $\boldsymbol{q}$ is introduced:

$$ {}^{g}\mathbf{x}' = {}^{g}\,\mathbf{H}({}^{c}\boldsymbol{x}')\mathbf{x}' $$

resulting in general image coordinates ${}^{g}\mathbf{x}'$ and using

$$ {}^{g}\mathbf{H}({}^{c}\boldsymbol{x}') = \begin{bmatrix} 1 & 0 & \Delta x'({}^{c}\boldsymbol{x}', \boldsymbol{q}) \\ 0 & 1 & \Delta y'({}^{c}\boldsymbol{x}', \boldsymbol{q}) \\ 0 & 0 & 1 \end{bmatrix} $$

where the terms $\Delta\mathbf{x}'({}^{c}\boldsymbol{x}', \boldsymbol{q})$ and $\Delta y'({}^{c}\boldsymbol{x}', \boldsymbol{q})$ are local corrections depending on $\boldsymbol{q}$ usually defining polynomials. Then real cameras showing distortions that do not preserve straight lines, most notably radial distortions of the lens system, can also be modeled.

The three dimensional object structure can be inferred from two images taken from two different places. For this purpose corresponding points P'_i and P''_i in the two images are measured. For a perfect orientation of the cameras the two corresponding rays $P'O'$ and $P''O''$ from the image points through the projection centers would intersect in the object point P. This is the so-called coplanarity constraint, since the corresponding rays of an oriented image pair are coplanar (Fig. 2).

An explicit expression for the coplanarity constraint for the relative orientation of two cameras is given by

$$ \mathbf{x}'^{t}\,\mathbf{F}\mathbf{x}'' = 0\,. \tag{4} $$

For a derivation from the projections of the two cameras according to Eq. 3 see [1]. The 3×3 fundamental matrix $\mathbf{F}$ is determined by seven independent parameters, as it is homogeneous and singular. Therefore, only seven corresponding points are necessary to determine its elements.

The point P'' in the second image corresponding with a point P' in the first image is located on a straight line. This line, called the epipolar line, is very helpful when searching for corresponding so-called homologous points. The underlying geometry is the epipolar geometry (Fig. 3). The epipolar plane ε (P) defined by the projection centers O' and O'' and the object point P intersects the image planes ε' and ε'' at the epipolar lines $l'(P)$ and $l''(P)$. The epipolar lines of all object points intersect at the epipoles E' and E''. These entities can be determined using the projection matrices or the fundamental matrix. Most importantly, due to the coplanarity constraint and the incidence of image points and epipolar lines $\mathbf{x}'^{t}\,\mathbf{l}' = 0$ and $\mathbf{x}''^{t}\,\mathbf{l}'' = 0$, the epipolar lines are given by

$$ \mathbf{l}' = \mathbf{F}\mathbf{x}'' \quad \mathbf{l}'' = \mathbf{F}^{t}\mathbf{x}'\,. $$

The relative orientation of three images gives constraints on all image coordinates involved. As the previous treatment of epipolar geometry shows, this is not the case for the image pair which only gives constraints in one direction. Therefore, it is useful to investigate the geometry of the image triplet expressed by the so-called trifocal tensor. It can be used to predict points and lines given in two images in the third one. The prediction of a line l' in the first image from given lines l'' and l''' in the other images can be obtained from

$$ \mathbf{l}' = \begin{bmatrix} \mathbf{l}''^{t}\mathbf{T}_1\mathbf{l}''' \\ \mathbf{l}''^{t}\mathbf{T}_2\mathbf{l}''' \\ \mathbf{l}''^{t}\mathbf{T}_3\mathbf{l}''' \end{bmatrix} \tag{5} $$

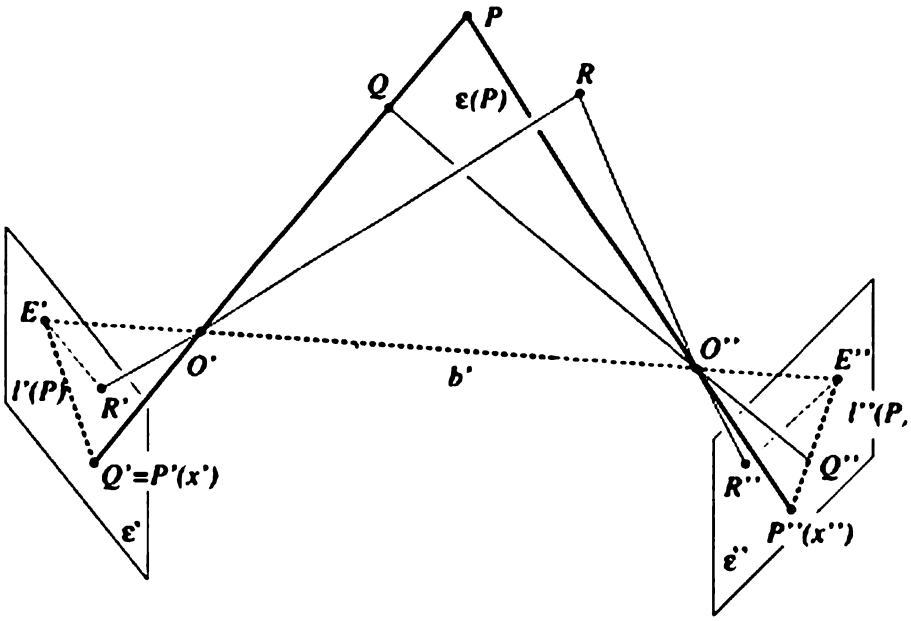
where $\mathbf{T}_i$ are the three 3×3 trifocal matrices stacked in the trifocal tensor $\mathbf{T}$. In simplified notation, (5) is expressed as [6,7]

$$ \mathbf{l}' = \mathbf{T}(\mathbf{l}'', \mathbf{l}''')\,. $$

Similarly, points can be predicted.

There exist relations between projection matrices, fundamental matrices and trifocal tensors. Owing to spatial limitations these derivations are not given here.

The three-dimensional position of object points can be determined by intersecting the rays defined by their image points and the corresponding projection centers. This is called photogrammetric triangulation and frequently based on multiple images. As the imaging rays of a camera form a bundle and the images overlappingly cover the object



Photogrammetric Methods, Figure 3 Elements of the epipolar geometry: epipolar plane $\varepsilon(P)$ through $O'O''P$, with the epipoles E' and E'' as images of the other projection center (O' or O''), the epipolar lines $l'(P)$ and $l''(P)$ which are the intersections of the epipolar plane $\varepsilon(P)$ and the image planes ε' and ε''. The epipolar planes build a pencil of planes with the base line $b' = O'O''$ as axis, e. g., induced by a different point R. Therefore the epipolar lines also build a pencil of lines with the epipoles as carrier. Observe, P' does not allow inference of where P sits on the projecting line. Point Q, also mapping to P', however, has a different image Q'' inducing the epipolar line $l''(P) = (E''Q''P'')$ in the other image [1]

space, this process is also called bundle block triangulation. It is usually formulated as a least-squares adjustment problem often termed bundle adjustment with, e. g., image coordinates and object space coordinates of control points as observations, and orientation parameters of the cameras and object space coordinates as unknowns. For instance, the collinearity equations, (3), could serve as observation equations in a Gauss–Markov adjustment model.

Key Applications

Photogrammetric 3D reconstructions are applied in various application such as acquisition of geoinformation, topographic mapping and terrain model generation using airborne imagery, and all kinds of close-range 3D reconstructions, e. g., in architecture, archeology, and engineering, as well as many applications in computer vision.

Future Directions

Recent developments include mathematical modeling and application of nonpinhole and uncalibrated cameras in real-time environments requiring direct (explicit) solutions for unknown parameters.

Cross References

- ► Data Acquisition, Automation
- ► Laser Scanning
- ► Photogrammetric Applications
- ► Photogrammetric Products

Recommended Reading

1. McGlone, J.C. (ed.): Manual of Photogrammety, 5th edn. American Society for Photogrammetry and Remote Sensing. Bethesda, Maryland (2004)
2. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An Invitation to 3-D Vision. Springer Verlag, New York (2004)
3. Mikhail, E.M., Bethel, J.S., McGlone, J.C.: Introduction to Modern Photogrammetry. Wiley, New York (2001)
4. Swaminathan, R., Grossberg, M.D., Nayar, S.K.: A perspective on distortions. Proc. Comput. Vision Pattern Recog. 2:594–601 (2003)
5. Abdel-Aziz, Y.I., Karara, H.M.: Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Proc Symp. Close-range Photogrammetry, pp. 1–18, Am. Soc. Photogrammetry (1971)
6. Faugeras, O.D., Luong, Q.T., Papadopoulo, T.: The Geometry of Multiple Images. MIT Press, Cambridge, MA (2001)
7. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, New York (2003)

Photogrammetric Products

J. CHRIS MCGLONE
SAIC, Chantilly, VA, USA

Synonyms

Map data; Cartographic data; Ortho-image; DEM; Digital elevation model; Elevation reference surface (datum); Digital surface model; TIN; Triangulated irregular network; Georectified; Ortho-mosaic; Oblique images; Root-mean-square error; RMS error; National map accuracy standard; 3D models; Photo-textured

Definition

A photogrammetric product is a representation of aspects of a scene derived from imagery of the scene. The representation may be geometric and include point coordinates, object geometry or measurements, or other attributes derivable from image geometry. In some cases, qualitative object properties may be added onto the basic geometric data.

Historical Background

Traditionally, photogrammetric products meant hardcopy maps depicting elevation as contours and features as lines. With the advent of digital softcopy photogrammetry for production and the widespread adoption of GIS to utilize cartographic data, emphasis has shifted almost exclusively to products in digital form. The increasing availability of digital imaging sensors has accelerated this trend. Indeed, the most rapidly growing types of photogrammetric products involve digital imagery, geo-located and processed for various GIS and consumer applications and delivered over the Internet.

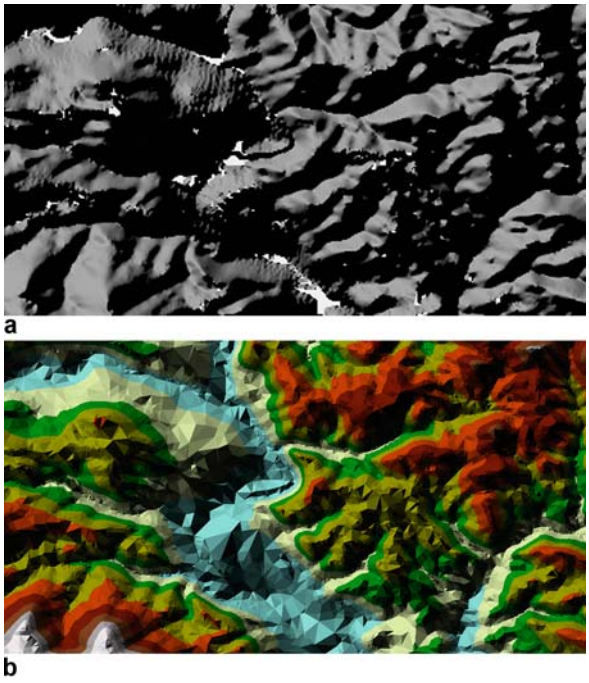
Scientific Fundamentals

The basis of a GIS is its geospatial information; photogrammetry is unique in its ability to provide accurate spatial data with high information content over wide areas. The orthoimage is a prime example; when used as a GIS base map, it provides a foundation for positional information as well as context for the display and interpretation of other data.

Elevation Products

Elevation products [1] represent the elevation of the earth's surface. Both raster and triangulated irregular network (TIN) representations are used, with the choice depending on the particular application.

Raster representations [1,2] are the most common, since they can be displayed and manipulated using standard image processing software and hardware. DEMs can be defined relative to any coordinate system, either projected coordinate systems such as UTM or directly in latitude-longitude. DEMs are described by their resolution or *post spacing*, the distance between adjacent elevation samples. For instance, USGS DEMs are usually described as 30-meter or 10-meter DEMs (for those in UTM), while NGA Digital Terrain Elevation Data (DTED) comes with 3-arc-second or 1-arc-second (latitude-longitude) post spacings. European DEM products include, from the UK, Land-Form PROFILE® Plus (2 m grid, 0.5 m RMSE for urban and flood plain areas up to 10 m grid with 2.5 m RMSE for mountain and moorland areas), from Germany, ATKIS DGM5 or DGM 25 with 5 m (not available everywhere) or 25 m grid spacing, and from France, BD ALTI® with 50 m grid spacing.



Photogrammetric Products, Figure 1 **a** Elevation raster, shown in hill-shaded form. **b** TIN version of the raster

Another important property of DEM products is the elevation reference surface, or datum. DEMs in the past were referenced to local height datums relative to sea level, but today are usually referenced either to a global geoid model (e. g., GEOID99) or to a reference ellipsoid (e. g., WGS84).

TIN representations [1,3] consist of a set of irregularly-distributed points connected by edges to form a surface consisting of connected triangles. TINs are typically more efficient than rasters in terms of the storage space required for an equivalent level of detail or accuracy, since more points can be concentrated in complex areas and fewer points used in flat areas, although three coordinates must be stored for a TIN point versus only the Z coordinate for raster representations. TIN points can be placed at the edges of breaks or in the bottoms of depressions in the terrain, whereas a raster's fixed sampling interval may not capture such terrain detail. In some cases rasters are augmented by breaklines which depict abrupt changes in surface slope.

TINs are more complicated to display than rasters since 3D graphics are required instead of simple raster displays.

TINs work well for graphics and simulation applications since current graphics cards are highly optimized to deal with sets of triangles. Exploitation is also more complicated, since determining the elevation at a given X,Y coordinate requires first identifying the triangular face of the TIN containing it, then interpolating the elevation from the three vertices of the face.

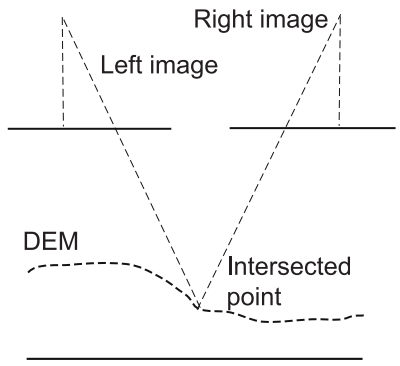
Contours (lines of equal elevation) [1,3,4] were formerly the standard elevation representation, due to ease of visual interpretation and their suitability for photogrammetric extraction. However, they are now secondary products derived from rasters or TINs as required.

Most elevation models depict the earth's surface as it would appear without buildings or vegetation and are referred to as digital elevation models (**DEM**) or digital terrain models (**DTM**). Constructing a DEM requires manual interpretation of the scene to remove non-terrain objects and to estimate the terrain elevation. Current automated processes such as automated stereo correlation or 3D sensors such as LiDAR or IFSAR represent the first (reflective) surface, containing buildings and the tops of trees or vegetation. These are referred to as digital surface models (**DSM**) and may be used for orthoimage production. Automated editing methods are somewhat successful in reducing DSMs to DEMs, but some manual editing is still required.

The majority of DEMs are currently produced by photogrammetric methods, although 3D sensors such as LiDAR are being rapidly adopted. To produce a DEM, the operator views the scene in stereo and places a 3D measuring dot superimposed on the model on the ground at the desired post spacing. Alternatively, the operator may capture 3D points at representative locations on the terrain and thereby generate a TIN from which a DEM can be interpolated if required. Automated stereo methods replace the operator by performing the stereo matching using image correlation techniques.

DEMs may be distributed in image formats, such as geotiff, or in special data formats such as USGS DEM or NGA DTED. DEM specifications typically specify the Z RMS error against some number of independently measured elevation points. Common DEM errors include noise spikes or pits due to measurement or processing errors. There may also be systematic offsets due to operator biases or caused by automated processes measuring the tops of vegetation instead of the ground surface.

Image Products Before an image can be used in a GIS, there must be some means to relate the locations of objects within the image to their locations in the world. This geometric relationship between a pixel in the image and a point on the ground is embodied in the *sensor model*, sometimes referred to as the image metadata. The sensor model includes:

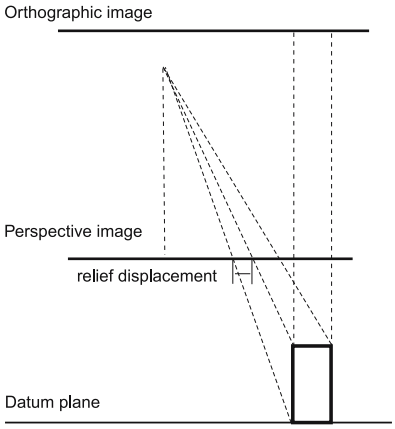


Photogrammetric Products, Figure 2 Determination of the 3D position of a point by the intersection of rays from two images or by the intersection of a ray from either image with a DEM

- the location of the perspective center of the sensor in world coordinates and the angular orientation of the camera with respect to the world coordinate system. This is sometimes referred to as the *exterior orientation*, since it places the sensor within an exterior reference system. For non-frame sensors, these parameters may be expressed as functions of time to model the path of the aircraft or the orbit of the satellite carrying the sensor.
- the *interior orientation*, the geometric parameters of the sensor itself. This includes the principal distance (the focal length for images at infinity) and the specification of origin, orientation, and scale of the image coordinate system.
- a set of equations, based on the principles of perspective geometry and using the parameters of exterior and interior orientation, which relates a point in the image to a point in the world.

Given the sensor model, one can model the path of a ray of light from a point in the world through the perspective center of the sensor and onto the imaging plane to calculate its image coordinates; alternatively, one can use the sensor model and the image coordinates to calculate the ray in space passing through the object in the world. Note that a single image can specify only the direction in space to an object: to calculate the 3D position of an object we must intersect rays from two or more sensors or else have external knowledge of the scene geometry, such as a digital elevation model, and intersect the ray with that surface to determine a 3D position.

Very few GIS include the capability to deal with the variety of sensor model types currently in use. Additionally, perspective effects present in unprocessed images make their combination with other types of data problematic. There-



Photogrammetric Products, Figure 3 Orthographic and perspective projections, showing relief displacement due to the height of the object

fore, images are usually processed to transform them into a more easily exploited form, both in terms of appearance and sensor model.

Rectified or **geo-rectified** imagery [2] is produced by reprojecting the image to a reference surface. Rectified frame photos were widely used in the past, since the film could be transformed into an equivalent vertical photograph using an analog rectifier. Digitally geo-rectified images are reprojected to a reference plane or, in the case of satellite imagery, the ellipsoid surface. This removes perspective effects, but does not correct for displacements due to differing elevations across the scene.

Orthorectified images (**orthoimages**) [2,4] are produced by transforming the original image into an orthographic projection. In an orthographic projection the projection direction is perpendicular to the datum plane, as in a map (Fig. 3), whereas in a perspective image objects above the datum plane are displaced proportional to their height (relief displacement). Since objects in an orthographic projection are shown at their true map locations, orthoimages are often used as base layers in GIS databases. Common examples of orthoimage products include the U.S. Geological Survey's Digital Ortho Quads, the UK Ordnance Survey Mastermap® Imagery Layer (0.25 m), the German ATKIS DOP (0.1–0.4 m), and the French BD Ortho® (0.5 m). One of the most visible current applications of orthoimagery is as a base for systems such as Google Earth, Microsoft Virtual Earth, and NASA WorldWind.

Orthoimage production requires the 3D coordinates of each point in the image, usually obtained by intersecting the image rays with a DEM of the scene. For each X,Y pixel location in the final orthoimage, the elevation is determined from the DEM and the coordinates are projected into the perspective image. The intensity value at that point in the orthoimage is then set to that of the perspective image. If multiple input images are available, the orthoimage intensity value can be determined as a combination of the various input images.

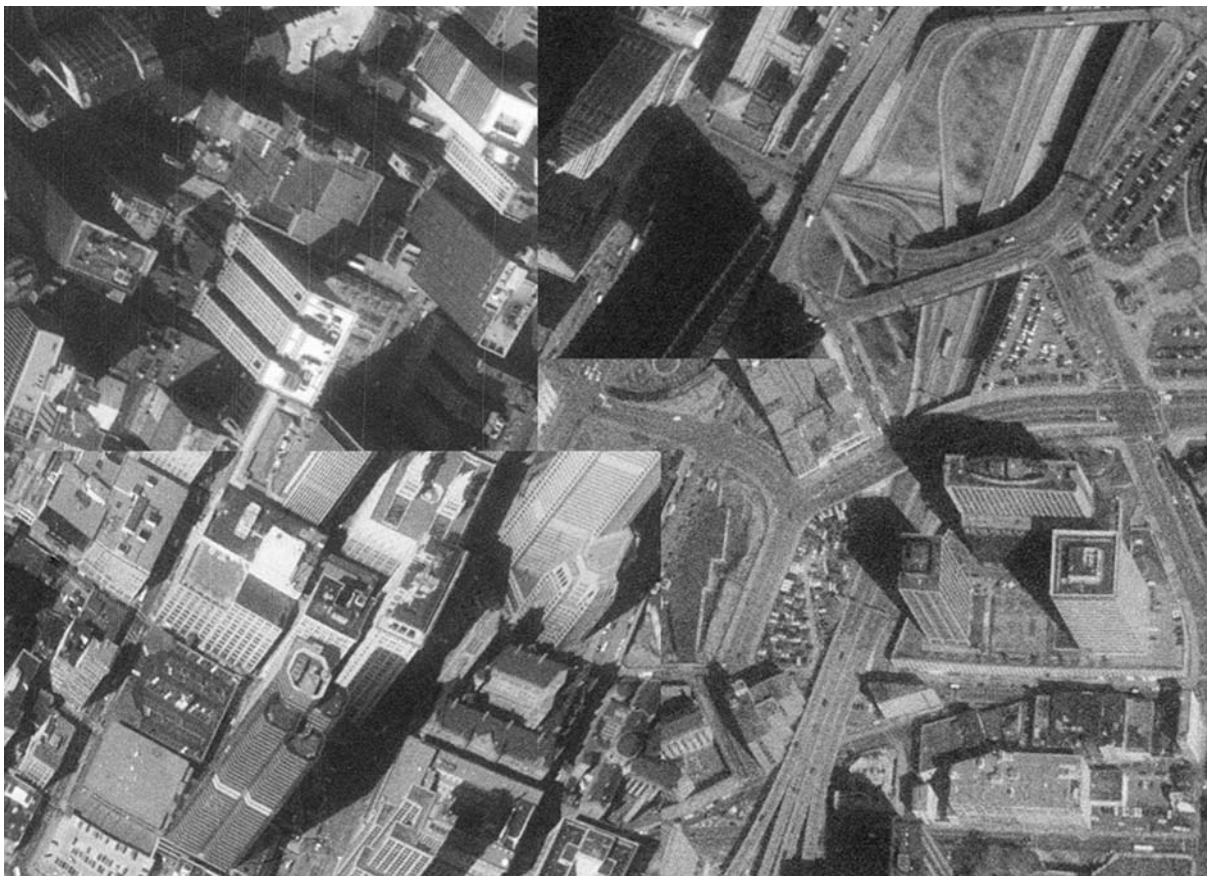
However, a digital elevation model contains the elevations of the terrain surface and not the elevations of structures on the terrain. Buildings in the scene will therefore appear to lean in the orthoimage (Fig. 4), due to the image perspective. Sometimes a digital surface model (DSM) is used, which contains the elevations of the terrain and objects on it. A DSM may be produced by photogrammetric methods or by direct 3D sensors such as LiDAR or IFSAR. Alternatively, 3D building models may be manually extracted and used in conjunction with the DEM to allow building roofs to be projected into their correct positions and occluded areas to be identified. Orthoimages produced in this manner are often referred to as **true orthoimages**. If multiple images are available, the area occluded by the building can be filled in from other viewpoints. To eliminate building shadows from the final image, shadows can be detected by comparing intensities among the images, or the shadow geometry may be predicted from the sun angle.

Most orthoimages are actually **orthomosaics** [2] produced from multiple images, permitting the coverage of large areas and the selection of the best image for any particular point. The images must be carefully blended for radiometry and color balance and the seam boundary between images must be carefully drawn to make it invisible.

Orthoimages are not suitable for all applications, since they show only building roofs and outlines which are hard to recognize from street level. Oblique aerial views (Fig. 5) show building facades and make building recognition and the determination of characteristics such as the number of floors much easier. Several companies now offer oblique aerial imagery covering sites from different angles, along with the associated sensor models and tools to enable their exploitation. The tools are designed to work either as plugins to GIS or to interoperate with GIS tools and allow measurements and positioning from the imagery. The positioning accuracy is typically limited by the precision of the navigation information.

Oriented image stereopairs may also be supplied, with their associated metadata, allowing exploitation within GIS using photogrammetric software designed to work within GIS packages. Several commercial satellite companies supply such stereopairs and support data, ready for exploitation by mapping companies. This is less common for aerial photography, since few users are equipped to do stereo extraction.

Distributing oriented imagery requires that the exploitation software implement the appropriate sensor model. Given the wide variety of sensor types currently in use, both aeri-



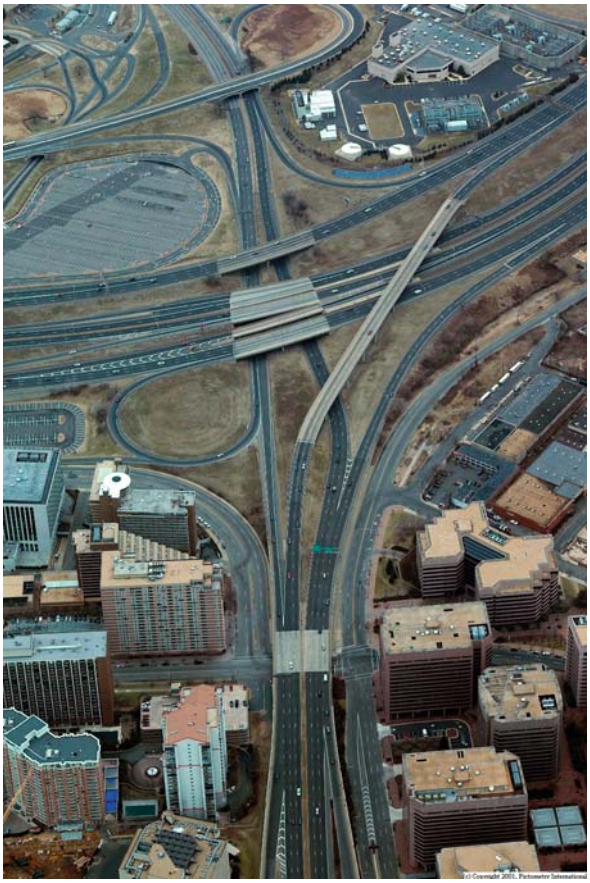
Photogrammetric Products, Figure 4 Apparent building lean in an orthomosaic, particularly evident at the mosaic seams. Notice that features at ground level are aligned

al and satellite, this poses a major implementation, verification, and maintenance burden. For this reason, there has been much work on "generic" or "replacement" sensor models, where the geometry of the sensor is modeled by a set of polynomials derived from the original physical model by the supplier of the imagery. Exploitation systems then only have to support the common polynomial model.

Image products may be produced in proprietary formats, but nearly all are available in standard formats. Orthoimages are widely distributed in geotiff format, which is based on the popular tiff image format and includes header tags which contain the coordinate and projection information. A newer standard is the JPEG2000 format, which uses the JPEG2000 image compression format along with header tags for geopositioning information. The US Dept. of Defense has defined the National Image Transmission Format (NITF), which allows for the inclusion of multiple images and their associated metadata as well as graphic overlay and text information within the same file structure.

Image product specifications have two main aspects, radiometric and geometric. Radiometric specifications are concerned with the appearance and interpretability of the image product. Good contrast and brightness are crucial; both qualities refer to the distribution of pixel values across the image histogram. Panchromatic (gray scale) images most often have 8 bits per pixel, meaning that they can represent 256 shades of gray. An image with good contrast will distribute the pixel values across nearly all 256 possible values, maximizing the visual information content. An image with good brightness level will have the values peaking near the middle of the histogram, instead of concentrating at one end or the other. Many digital sensors now collect data with 11 or 12 bits of radiometric resolution. This greater range must be mapped into the typical 8 bit display while preserving both fine detail and overall structure and contrast.

The color balance of the image is also important, both its relationship to the original colors in the scene (assuming that it is a true color image, as opposed to a false-color or



Photogrammetric Products, Figure 5 Oblique aerial image, supplied with exterior orientation parameters to allow measurement and positioning. Courtesy of Pictometry, Inc

infrared image) and its appearance on the computer monitor. Tasks requiring precise image interpretation based on color require the calibration of the monitor to color standards as well as correction for the color response of the sensor.

Improvements in digital sensors have greatly reduced the occurrence of image artifacts such as streaks or blooming due to overexposure, but there may still be issues related to the acquisition, such as haze, cloud cover, or bright spots resulting from the relative alignment of the sun and sensor.

The geometric accuracy of an image product is specified in terms of the error computed by comparing identifiable image points with independent coordinate measurements. These measurements will be in X and Y for orthoimage products, or in all three coordinates for oriented stereopairs. The accuracy is typically specified in terms of the root-mean-square (RMS) error over a given number of well-distributed check points. For instance, the USGS specification for Digital Ortho Quads requires that they meet National Map Accuracy Standards at 1:24,000 scale: 90 percent of the well-defined points tested must fall within 40 feet.

Vector Feature Products Vector products [3,4] may be considered "line drawings" of objects or areas in the scene. They are most often used to describe man-made objects such as roads or buildings, or to delineate areas. Vectors features have three aspects: geometry, topology, and attribution.

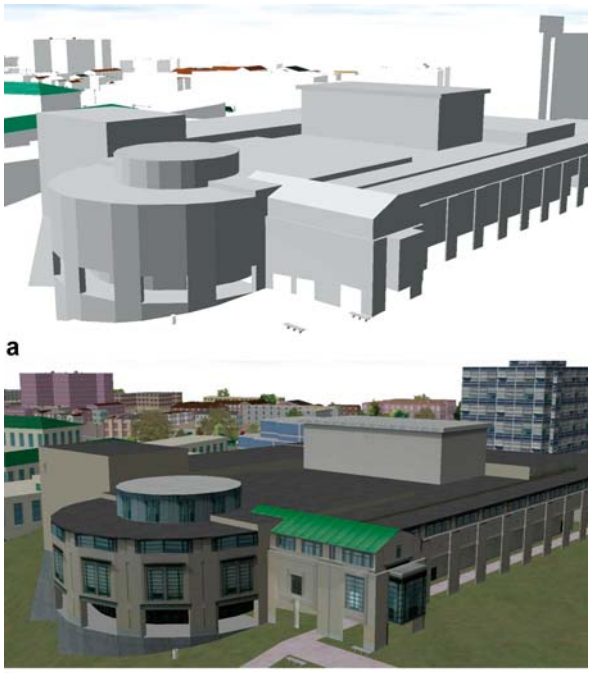
The geometry of vector features is usually expressed in terms of the coordinates of individual points, although sometimes the feature will reference a defined geometric form such as a circular arc, spline, or rectangle. Features may be points, lines, or polygons. Lines may be "polylines" consisting of multiple connected lines. A 2D feature has only X and Y coordinates; if the Z coordinate is included for each point the feature is often called "2.5D," since non-horizontal faces of 3D objects such as buildings are not explicitly represented. A true 3D representation contains information on object faces, such as the direction of the outward-facing normal vector (cf. below).

A topological representation allows reasoning over spatial relationships between objects, such as "adjacent to," "inside," or "connected to." For instance, for a given line, connected lines and adjacent faces can be easily determined. The elements of a topological representation are nodes, points, lines, and faces. Nodes are connections between lines, as opposed to points which only indicate position. Lines consist of ordered sets of points, and connect to other lines only at nodes. A line has a left and right side and may be the border of a face. Faces may contain other faces and also be contained within larger faces.

Topological encoding enables reasoning about the properties of a scene. For instance, a topologically-encoded road network allows reasoning about routes, through the connectivity information, or about the access of parcels of land represented as faces to the road network, since the neighboring face of each line is defined. The collection of topological information also allows collected data to be checked as it is added.

Vector features are usually collected with *attribution* which stores visible or inferred properties of the feature. Road features may contain attributes which indicate the type of surface (concrete, asphalt, gravel), the number of traffic lanes, or the type of road; buildings attributes may include an inferred use (residential, commercial, factory) or type of construction (brick, frame).

Specifications for vector data typically concern spatial accuracy, the types of features to be captured, and the degree of detail captured. Some examples of current vector datasets include USGS Digital Line Graph, UK OS



Photogrammetric Products, Figure 6 **a** 3D building model without phototexture. **b** Phototextured model

Mastermap® Topography Layer, German ATKIS Basis-DLM and DLM50, and French BD TOPO®.

3D Models An increasingly-common product is 3D models [3], used for both GIS and visualization applications. Most photogrammetric 3D models are collected as wireframes, which are well suited for buildings composed of planar surfaces, although some applications use combinations of 3D geometric primitives. If 2D building footprints are available, they may be extruded upward to the measured building height to produce a 3D model, but in most instances 3D models are extracted by measuring corner points.

The photogrammetric extraction of 3D models is extremely labor intensive. While some semi-automated systems are being used in production, the problem for general buildings is so complicated that efforts at fully-automated extraction have not yet been successful. The main cost driver is the level of detail which must be represented. For telecommunications applications concerned with signal propagation, fairly coarse models are acceptable, while planning or security applications often require extremely detailed models so that calculated lines of sight are accurate or building appearances are realistic.

Building models are often *phototextured,* (have imagery applied to the faces) for a more realistic appearance. The simplest texturing is done by applying repeating generic wall patterns to building faces. When more realism is required, actual imagery of the building is used. For purely visual applications, building phototextures can often be substituted for detailed geometry—i. e., instead of collecting detailed façade structure, an image of the façade applied to the model will give the impression of the geometry without the expense of detailed extraction. Aerial imagery seldom works well for texturing, since the viewing angle is nearly parallel to the building surfaces and results in a "smeared" appearance. Oblique aerial imagery or ground imagery is usually much more satisfactory.

3D formats for GIS use are problematic. Most common 3D formats are designed for visualization purposes, supporting geometry and textures but not attribution. The most common example is OpenFlight® by Multigen-Paradigm. Several vendors do offer proprietary solutions to allow attribute query on 3D representations.

Key Applications

Photogrammetric products provide the location basis for nearly all GIS applications, either as the primary source of the information or as the context and framework into which data is conflated and utilized.

Future Directions

The trend to 3D products, especially building models, can be expected to continue as more consumer applications based on such products are introduced. The visual aspects will become more important, especially for consumer applications where absolute cartographic accuracy is less important. Photogrammetric products are just one component of increasingly complex and multi-faceted databases, which may include addresses, demographic information, or commercial and advertising content.

There will be an ongoing competition between the automation of feature extraction processes and its outsourcing to countries with lower-cost labor. For many areas, the task will change from feature extraction to maintenance, update, and enhancement of existing GIS systems. Licensing of portions of large-area datasets maintained by one vendor, such as road or address databases, will become more common than the production of data for a single user over a specific area.

Product distribution will occur through a growing number of channels, especially outside the traditional GIS/mapping industry. Products will increasingly be embedded in consumer applications and targeted to non-cartographic users and applications, for instance, 3D building models delivered to cell phones for location-based services.

Cross References

- ► Photogrammetric Methods
- ► Visualizing Constraint Data

Recommended Reading

1. Maune, D. (ed.): Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd edn. American Society for Photogrammetry and Remote Sensing, Bethesda (2006)
2. Miller, S.: Photogrammetric Products. In: McGlone, J.C. (ed.) Manual of Photogrammetry, 5th edn. American Society for Photogrammetry and Remote Sensing, Bethesda (2004)
3. Mikhail, E.M., Bethel, J., McGlone, J.C.: Introduction to Modern Photogrammetry, John Wiley and Sons, New York (2001)
4. Wolf, P.R., Dewitt, B.A.: Elements of Photogrammetry with Applications in GIS, McGraw-Hill, New York (2000)

Photogrammetric Sensors

MICHAEL CRAMER
Institute for Photogrammetry (ifp),
Stuttgart University, Stuttgart, Germany

Synonyms

Photogrammetric images; Photogrammetric cameras; Aerial imagery; Image acquisition; Air borne sensors; Remote sensing

Definition

Images are the main data source of photogrammetric data processing. Hence the sensors used for data acquisition are an elementary part of the photogrammetric processing chain. In general, images are taken by satellite, airborne, or terrestrial sensors for photogrammetric applications such as object and terrain modeling and acquisition of topographic data. This entry summarizes the state-of-the-art and focuses on the strong trend towards digital image recording.

Historical Background

The idea of using photographs for the reconstruction of the imaged objects was born almost at the same time as the invention of photography. Starting in the mid of the 19th century object coordinates were estimated from two dimensional images based on the fundamental equations of image geometry. The term "photogrammetry" appeared for the first time in 1867 [1]. In the beginning photogrammetric reconstructions were limited to terrestrial applications, although first experiments to obtain imagery from the air were already performed with balloons or kites long before 1900. The success of those attempts has been limited due to the restricted maneuverability of the camera carrier. The situation rapidly changed with the advent of airships and airplanes. Since then photogrammetry using aerial photography has been established as the preferred method for the mapping of large areas.

Pushed by the growing need for airborne images the technical design of the cameras was continuously refined. The first prototype of an aerial camera for serial photography was already presented in 1915 [2]. Further on, the quality of camera lenses and the image format was continuously increased to obtain larger terrain coverage per image. The cameras were initially used hand-held, which was less optimal for the layout of the image block formed from several overlapping images. Thus, later airborne sensors were fixed to aircrafts with special camera holes in their body to realize a vertical viewing direction. High performance and geometrical stable roll film material substituted glass plates. Refined and efficient techniques for image data recording were introduced. They comprise high quality optical systems with extremely high resolving power, forward motion compensation, stabilized platform mount, photo flight navigation and aircraft guidance, as well as direct measurement of sensor's exterior orientation during flights.

Scientific Fundamentals

The acquisition of imagery is based on the principles of photography. Photography is a passive method, i. e. the energy, reflected from the object is recorded by photo sensitive material or elements. Until recently this was exclusively done using analogue films. They are now increasingly replaced by digital sensor elements. Consequently, today in operational photogrammetric environments analogue as well as digital sensors are employed.

The benefits of direct digital image recording in comparison to the former digitization of analogue imagery via scanning are obvious: There are cost and time savings because analogue films and film development is not necessary any longer. The time consuming film scanning is dispensable. Besides this, digital recorded images provide better radiometric quality. This positively influences the later automated point measurements in photogrammetric processing. Digital sensors allow for a parallel acquisition of pan-chromatic and multi-spectral image data opening up new fields of application.

Traditionally, photogrammetry is concerned with three-dimensional object reconstruction from two-dimensional images. Similar to human stereo vision photogrammetric object reconstruction is based on two or more different images from the same object with certain image overlap. This need of sufficient image overlap results in spe-

cific image block structures. In terrestrial or close range applications the images are mostly taken all around the object of interest realizing convergent viewing directions. In airborne applications the image acquisition is following a pre-planned regular flight pattern. The images themselves are arranged in flight lines with overlaps within the flight line and between neighbouring lines. For satellite imagery the flight path is determined by the satellite's orbit. Many satellites are able to steer their imaging sensor in cross-track and/or along-track orbit direction to obtain stereo image coverage.

Since photogrammetry focuses on the precise geometric reconstruction, the design of imaging sensors has to follow certain requirements. Within three-dimensional object reconstruction the correct reconstruction of imaging rays is essential, i. e. the determination of the sensor's interior and exterior orientation. In order to obtain the interior orientation of the sensor, a calibration is performed. Until recently such calibration was mainly done in laboratories. Today there is a clear trend towards alternative calibration methods [3].

Particularly, the digital imaging sensors become more complex and heterogeneous. Some of them are using more than one optical component and they are often equipped with additional components such as navigation sensors. Calibration is shifted towards a more system oriented approach. From this view point in-situ calibration provides a powerful tool to calibrate and validate such digital sensors. It is already well established for geometrically less stable imaging sensors, e. g. for terrestrial close-range applications. Here, in contrast to the stable geometry of the traditional mapping cameras changes in sensor geometry over time prevent an a priori laboratory based calibration.

In the remainder the focus is on the sensors used for airborne photogrammetric applications, as this is the by far largest area of photogrammetry. Typically, large image formats are employed in order to guarantee efficient data acquisition. The available image format directly influences the effort to cover a certain area with imagery. Therefore, traditional analogue mapping cameras have been designed for very large format films with standard formats of about $23 \times 23\,\mathrm{cm}^2$. Typically focal lengths of 30 cm (normal angle, yields a field of view (FOV) from corner to corner of approximately 60 deg), 15 cm (wide-angle, FOV approx. 95 deg) and 8 cm (super wide angle, FOV approx. 125 deg) are used to adapt to different application scenarios. Analogue mapping cameras have been manufactured by different system suppliers, while the majority of analogue aerial imagery is taken by only two different mapping cameras, namely the Intergraph Z/I (formerly Zeiss) RMK-Top and the Leica Geosystems (formerly LH-Systems, Wild) RC30 series and their predecessors.

Both cameras are very similar, which in the past has pushed the development of measuring and data evaluation tools, independently from the imaging sensor itself. The major part of the mapping system is the camera body with the camera lens cone including the shutter and several lenses. In addition to its very high resolution the optical part has to fulfil geometric requirements leading to measurement accuracies in the range of a few micrometers. Differently from consumer cameras the mapping camera has to preserve its stable geometry for a long period of time and under changing environmental conditions. The camera is fixed in the aircraft (mostly) using an active mount which isolates the camera from the aircraft vibrations and additionally controls the attitude and heading of the camera. Due to this stabilization the airborne images are taken close to nadir viewing (i. e., horizontally) regularly. On the other hand, the active control of attitude variations minimizes the blurring effects caused by aircraft rotations during image recording. The remaining image blur which is due to the forward component of aircraft movement is typically compensated by shifting the film whilst image exposure. This is called forward motion compensation (FMC). As an example the Intergraph Z/I RMK Top15 is depicted in Table 1. The camera is fixed in the stabilized mount and on top of the camera body the removable film magazine can be seen.

Almost 2000 large format mapping cameras have been distributed all over the world over more than four decades. As of 2005 around 800 of them are still used in different kinds of operational applications. Nonetheless, the era of analogue imagery in photogrammetry comes to an end, similarly to the development in the consumer market. In May 2006 the camera manufacturer Intergraph Z/I announced that there are currently no plans to manufacture new RMK-Top systems. This proofs that digital sensors can compete with the analogue sensors. The official introduction of commercial digital airborne cameras started in 2000.

In order to obtain large formats in digital imaging, two different sensor designs are used. The first relies on a small number of digital CCD sensor lines which can be offered with a reasonable length. These lines are grouped perpendicular to the aircraft's flight direction. Full terrain coverage is obtained via the aircraft's motion. This line scanner concept is also named pushbroom scanning and is known from satellite imaging. In photogrammetry, the systems are often referred as three-line scanners, although typically more than three lines are used to obtain three panchromatic channels as well as four multi-spectral channels. All CCD lines provide the same number of pixels regularly. Thus pan-chromatic and multi-spectral images are obtained with the same geometric resolution. Since the individual physical placement of the lines within the focal

Photogrammetric Sensors, Table 1 The Intergraph Z/I RMK Top15 analogue mapping camera

Intergraph Z/I RMK Top15
– wide-angle lens, focal length 153 mm
– Angular field of view 93 deg (corner to corner)
– Aperture f/4 – f/22, continuously variable
– Exposure time 1/50s – 1/500s, continuously variable
– Remaining distortions $\leqq 3\mu$m
– Film length 150m with 0.1 mm film thickness
– Gyro stabilized mount with ± 5 deg in ω, ϕ, ± 6.5 deg in κ
– Weight ~165kg (including mount, magazine and control unit)

plane is different, each CCD line provides a different viewing direction.

Contrary to this, large format frame based sensors combine several individual camera heads, each one equipped with one or more CCD frame sensors. All these camera heads are fixed to one airborne platform. The smaller format images, taken by the separate individual camera heads, provide certain image overlaps. This allows for the generation of one synthetic large format image afterwards. Typically, the large format image is taken in the pan-chromatic channel. Additionally, multi-spectral channels are captured simultaneously by additional camera heads, but typically with less spatial resolution compared to the large format virtual pan image. High-resolution colour imagery is obtained from later processing, where the colour channels are combined with the high-resolution PAN images. In these cases, the original radiometric colour information will be more or less impacted, depending on the algorithm and ratio used. This process is termed pan-sharpening. Similar concepts are used in satellite imaging.

In general the frame based digital sensors try to transfer the classical concept of photogrammetric 2D image data processing from the analogue to the digital world, whereas the line scanning approach is following an alternative concept with its long image strips compared to the individual image frames from frame sensors.

Many of the digital airborne sensors are combined with additional sensors for the direct measurement of exterior orientation elements at the time of exposure. High performance integrated GPS/inertial systems are available to solve this task [4,5]. If the GPS/inertial exterior orientation elements are obtained with sufficient accuracy, the photogrammetric image orientation can be done without any additional ground control (so-called direct georeferencing) [6]. In case of pushbroom scanners the use of GPS/inertial components is inevitable. Due to their less stable image geometry (only 1-dimensional lines are recorded instead of 2-dimensional image frames) the additional GPS/inertial measurements are necessary to compensate for the image distortions caused by the sensor's movement during image data acquisition.

The following Table 2 briefly summarizes the main characteristics of three commercially available large format digital sensors. Two of them – namely DMC from Intergraph Z/I [7] and Ultracam-X from Microsoft (formerly Vexcel) [8] – are following the frame concept, whereas the Leica Geosystems (formerly LH-Systems) sensor ADS40 [9] is one representative of the pushbroom line scanners. Besides these, other digital sensors are available and new ones are emerging.

ADS40 and DMC were officially introduced to the market in 2000, whereas the Ultracam-D was presented in spring 2003. Already in 2006 modifications of Ultracam-D and ADS40, namely Ultracam-X and ADS40 (2nd generation) were presented to the photogrammetric community. Meanwhile, more than 100 systems altogether have been sold (status 2006), with the digital mapping sensor market quite equally distributed between all three sensors. System manufactures expect future system sales of ~15–20 individual systems per year. Restricted to the before mentioned three different sensors, this will result in an annual increase of 45–60 digital systems per year. Within another 5 years period the number of available large format digital sensors will be about 325–400 at least.

In addition to that, other camera formats with small to medium image sizes are also used in airborne applications [10,11]. Those systems are based on the frame sensor concept and now (2007) provide images about 4000×5500 pix or 7000×5500 pix. They were usually not designed for the use in airborne photogrammetric environments originally and, therefore, cannot fully compete with the large format sensors in terms of accuracy, image quali-

Photogrammetric Sensors, Table 2 Overview on today's commercial large format digital airborne mapping cameras

Sensor	DMC	Ultracam-X	ADS40, 2nd generation
Manufacturer	Intergraph Z/I	Microsoft/Vexcel	Leica Geosystems
# camera heads	4 + 4	4 + 4	1
Focal length PAN	120 mm	100 mm	62.5 mm
MS	25 mm	33 mm	62.5 mm
Image size PAN	13824 × 7680 pix	14430 × 9420 pixel	12000 × 1 pix staggered (nadir view)
MS	3000 × 2000 pix	4992 × 3328 pixel	12000 × 1 pix
Pan sharpening (for colour imaging)	yes, applied	yes, applied	no (original spectral bands)
Physical pixel size	12 µm	7.2 µm	6.5 µm
Field of View			
across-track	69 deg (PAN)	55 deg (PAN)	64 deg
along-track	42 deg (PAN)	37 deg (PAN)	
Optics	Carl-Zeiss Jena, system specific	Linos/Vexcel, system specific	Leica, system specific

ty and efficient coverage, especially for large area projects. Nevertheless, due to their reduced physical size and higher flexibility these systems can be accommodated in small aircrafts or even unmanned airborne vehicles (UAV). They are well suited for local use, e. g., repeated flights for monitoring or for quick response for disasters. It is interesting to note, that the number of small to medium format digital airborne cameras is by far exceeding the number of large format cameras today. In many cases they are not only used stand-alone but integrated with other sensors like laser scanners and direct georeferencing components.

Key Applications

Photogrammetry in its original sense is defined as the discipline of the quantitative analysis of photographs. The focus is on the reconstruction of three dimensional object geometry from two dimensional images. Traditionally the main field was in mapping.

The key applications are highly correlated with the source of sensor data and are divided in satellite, airborne, and terrestrial / close range applications. For satellite images the focus is on large area coverage and accuracies of 1 meter and less. For airborne photogrammetric tasks the areas covered are less extended, but typically geometric accuracies in the range of centimetres to few decimetres are required. The highest accuracies are required for close range applications, like highly detailed reconstruction of architectural sites or in industrial or medical environments. Especially for the later additional sensors besides cameras are used. Here maximum accuracies in the range of submillimeter to micrometer are realized.

Future Directions

As illustrated above, the world of photogrammetry is split at the moment: New digital sensors are used parallel to well established analogue cameras. Nevertheless, the future of photogrammetric data acquisition will be fully

digital – this is already almost the case for the processing of data.

Although the focus in the sections above was almost exclusively on airborne cameras, the world is more heterogeneous! Besides optical cameras (covering different image formats) other sensors based on laserscanning (LIDAR) or radar technology are used. In terrestrial and close range applications additional measuring devices such as stripe or pattern projectors are in use. In addition to pure geometric reconstruction multi-spectral data recording and analysis becomes more and more important. A clear trend towards multi-sensor systems is obvious. Traditional platform carriers such as aircrafts, helicopters and satellites are supplemented by new carriers like remotely controlled or autonomous flying model aircrafts, helicopters and airships for low flying altitudes. On the other hand, high altitude long endurance (HALE) UAVs allow for data recording from very high altitudes of more than 20 km. Pushed by various novel sensors and platforms new application areas, e. g. monitoring, disaster mapping, precision farming, real estate and tourism evolve. These applications have in many cases less stringent geometric accuracy requirements.

New applications require new concepts for system design and data processing. For example, for monitoring information is needed with a (very) high update rate but confined to a limited area of interest (i. e. daily monitoring of traffic flows in a city). As a consequence a fast and fully automatic processing of data is needed. Fully automatic image matching and feature extraction is highly desirable. In future these tasks might be solved directly on the sensor's chip. This ultimately could lead to intelligent sensor systems or sensor networks.

Cross References

- ► Photogrammetric Applications
- ► Photogrammetric Products

Recommended Reading

1. Meyer, R., Meydenbauer, A.: Baukunst in historischen Photographien. Fotokinoverlag, Leipzig (1985) ISBN: 3731100630, 259 pages
2. Albertz, J.: 90 Jahre Deutsche Gesellschaft für Photogrammetrie und Fernerkundung e.V. Photogrammetrie-Fernerkundung-Geoinformation PFG 5(1999), pp. 293–349 (1985)
3. Cramer, M.: EuroSDR network on digital camera calibration. IAPRS, Vol. XXXV, Part B, Istanbul, digitally on CD-Rom, (2004) 6 pages
4. Mostafa, M., Hutton, J., Reid, B.: GPS/IMU products – the Applanix approach. Fritsch/Spiller (eds.) Photogrammetric Week 2001, pp. 63–83, Wichmann Verlag, Heidelberg, Germany (2001)
5. Kremer, J.: CCNS and AEROcontrol: Products for efficient photogrammetric data collection. Fritsch/Spiller (eds.) Photogrammetric Week 2001, Wichmann Verlag, Heidelberg, Germany, pp. 85–92. (2001)
6. Schwarz, K.-P., Chapman, M.A., Cannon, M.E., Gong, P.: An integrated INS/GPS approach to the georeferencing of remotely sensed data. Photogr. Eng. Remote Sens. PE & RS, **59**(11), 1667–1674 (1993)
7. Hinz, A., Dörstel, C., Heier, H.: Digital Modular Camera: System concept and data processing workflow. IAPRS, vol. XXXIII, part B1, Amsterdam, pp. 164–171, digitally on CD-Rom. (2000)
8. Leberl, F., Gruber, M., Ponticelli, Bernoegger, S., Perko, R.: The Ultracam large format aerial digital camera system. Proceedings ASPRS 2003 annual conference, Anchorage, Alaska, May 2003, digitally on CD-Rom (2003) 7 pages
9. Sandau, R., Braunecker, B., Driescher, H., Eckart, A., Hilbert, S., Hutton, J., Kirchhofer, W., Lithopoulos, E., Reulke, R., Wicki, S.: Design principles of the LH Systems ADS40 airborne digital sensor. IAPRS, vol. XXXIII, part B1, Amsterdam, pp. 258–265, digitally on CD-Rom (2000)
10. Mostafa, M.: Design and performance of the DSS. Fritsch (ed.) Photogrammetric Week 2003, Wichmann Verlag, Heidelberg, Germany, pp. 77–87, 2003. Also published in Proceedings International Workshop on Theory, Technology and Realities of Inertial/GPS Sensor Orientation 22.–23. September 2003, Castelldefels, Spain, digitally on CD-Rom (2003)
11. Cramer, M.: Performance of medium format aerial sensor systems, IAPRS, vol. XXXV, part B, Istanbul, digitally on CD-Rom (2004) 6 pages

Photogrammetry

- ► Data Acquisition, Automation
- ► Intergraph: Real Time Operational Geospatial Applications
- ► Photogrammetric Applications

P

Photo-Textured

- ► Photogrammetric Products

Pixel

- ► Spatial Data Transfer Standard (SDTS)

Pixel-Based Prediction

- ► Bayesian Spatial Regression for Multi-source Predictive Mapping

Pixel Size

- ► Spatial Resolution

Plane Sweep Algorithm

JORDAN WOOD[1], SANGHO KIM[2]
[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
[2] ESRI, Redlands, CA, USA

Synonyms

Sweep line algorithm; Spatial join

Definition

The plane sweep (or sweep line) algorithm is a basic computational geometry algorithm for finding intersecting line segments. The algorithm can run in $O(n \lg n)$ time, where n is the number of line segments. This algorithm can be altered to solve many related computational-geometry problems, such as finding intersecting polygons. In spatial databases, we are generally looking at the special case of finding the intersections of minimum bounding rectangles (MBR, see definitional entry).

Historical Background

The plane sweep algorithm, or sweep line algorithm, originated from line segment intersection problem in computational geometry field. The paper and thesis written by Michael Shamos in the middle of 1970 first addressed the computational geometry problems. Later the book [1] written by Preparata and Shamos in 1985 contributed to making people widely aware of the problems. The plane sweep algorithm is one of the main topics in the book, along with other subjects such as convex hull, Voronoi diagram, and all-line-intersections. The plane sweep algorithm has been actively studied since then and expanded rapidly with numerous journal articles. Many application areas have also started using the algorithm. In robotics and motion sensing, it is critical to detect when any two objects intersect for collision. In computer graphics, the ray shooting method requires determination of the intersection of ray with other objects. In spatial databases, the spatial join algorithm deploys the idea of plane sweep algorithm as described in this article.

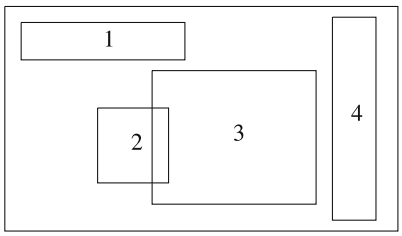
Scientific Fundamentals

The most common use of the plane sweep algorithm in spatial databases is to perform a spatial join. A spatial join is used to answer questions such as, "list all census blocks within 5 miles of an international airport." If we are dealing with a large dataset, such as a map of the US which contains all census blocks and all airports, this could be a very computationally intensive query. The naïve approach to this problem would be to create a set of 5 mile radius circles centered at the airports, then compare the polygon of each census block to see if it overlaps with each of the circles. This would be a very inefficient algorithm. Several different optimizations could be applied to this scheme. First, a spatial indexing scheme could be applied to organize the data so that only census blocks physically near an airport would be tested for overlap. This is an important optimization for the plane sweep algorithm, since it assumes that all the lines or polygons to be analyzed will fit in main memory. Also, a filter and refine strategy decreases the computation time for finding the intersection of complex objects, while increasing the number of objects which can be held in main memory at one time.

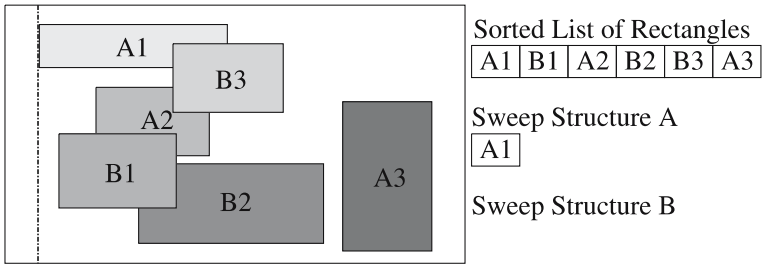
The problem of finding the intersecting rectangles from two sets of rectangles where exactly one rectangle comes from each set is interesting from a spatial database perspective, because the filter step of many spatial joins can be reduced to this problem. First, the geometric shapes are simplified to their minimum bounding rectangles (MBR). The filter step consists of finding the intersection of these MBRs. Then in the refine step, only those shapes whose MBRs have intersected are themselves tested for intersection. Orenstein [2] demonstrated the first use of the filter and refine technique for spatial joins using MBRs. A plane sweep algorithm is commonly used to find the intersection of the two sets.

Algorithm Description

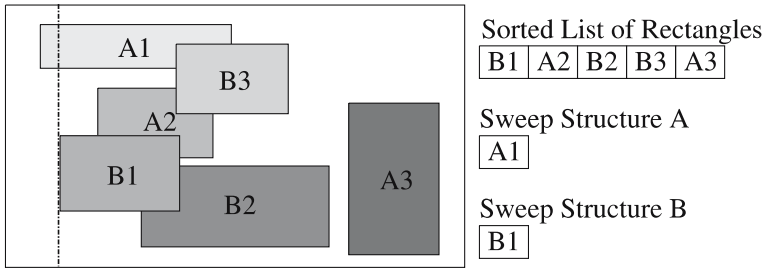
The plane sweep algorithm finds all overlapping MBRs from two sets. There are two phases to this process. First, the rectangles from both sets are sorted in increasing order based on their left sides. Figure 1 shows an example of this ordering. In the second phase, a vertical scan line is swept from the left to the right, stopping at each left rectangle side. All rectangles which are crossed by the scan line as it sweeps across the input are considered 'active'. Only active rectangles need to be tested for intersection.



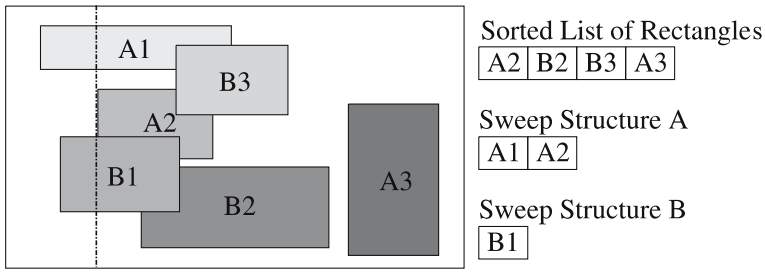
Plane Sweep Algorithm, Figure 1 A set of MBRs



Plane Sweep Algorithm, Figure 2 First iteration of the algorithm



Plane Sweep Algorithm, Figure 3 Second iteration of the algorithm



Plane Sweep Algorithm, Figure 4 Third iteration of the algorithm

Every time the sweep line stops, all inactive rectangles are removed, we test for intersection with the new rectangle, and we add the new rectangle to the active list. A data structure called a sweep structure, which can support the addition, removal, and intersection operations in a time efficient manner, should be used to store the active list. For example, a dynamic interval tree (which is a Cartesian tree) can be used.

Spatial joins require that we take rectangles which are partitioned into two sets, A and B, and find all the pairs of rectangles which overlap, where one rectangle is from set A and the other is from set B. We must therefore have a sweep structure for each of the sets. When the list of rectangles is sorted in the first phase of the algorithm, we must merge the two sets into a single list, while keeping track of which set the rectangle is a member of. As we consider each rectangle in the merged list, we test for intersection with the active rectangles which are members of the opposing set.

Example of Algorithm Execution

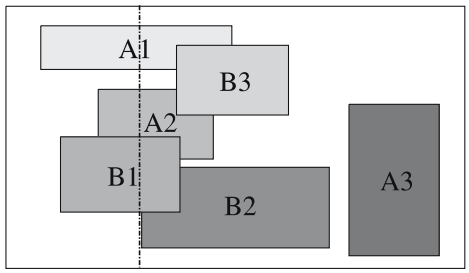
We will use the following example to illustrate the basic plane sweep algorithm used in the filter step of a spatial join. Set A has three members, as does set B. The rectangles are stored as the coordinates of their lower left and upper right coordinates (e. g., (A1.xl, A1.yl) and (A1.xu, A1.yu)). We use these values to determine whether a given pair of rectangles intersect, and whether a given rectangle is to the left of the sweep line. The sweep line is shown as a dashed line. When the sweep line reaches a rectangle from set A, we examine only the members of set B for intersection, and vice versa.

Figure 2 shows the first iteration of the algorithm. There are no elements currently in sweep structure B, so A1 cannot intersect with any of the members of B. At the end of this iteration, we remove A1 from the head of the sorted list.

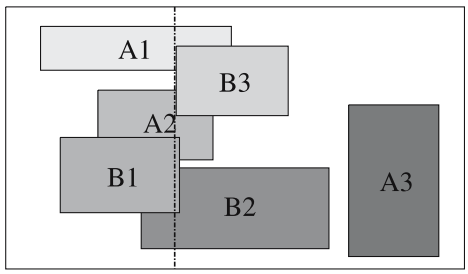
The second iteration is shown in Fig. 3. We compare the y-values of A1 and all the elements of sweep structure B (in this case just B1) to see if they intersect. They do not, and there are no intersecting pairs found in this iteration.

Figure 4 displays the third iteration. When we compare A2 with the one element in sweep structure B, we find that they do intersect. This iteration therefore produces the intersecting pair A2:B1.

The fourth iteration is shown in Fig. 5. We compare B2 to the elements of sweep structure A, and find no intersecting pairs.



Plane Sweep Algorithm, Figure 5 Fourth iteration of the algorithm



Plane Sweep Algorithm, Figure 6 Fifth iteration of the algorithm

The fifth iteration is shown in Fig. 6. We have removed B1 from sweep structure B because it is no longer crossed by the sweep line. This is not strictly necessary, but it keeps the sweep structures from continually growing.We compare B3 with the elements of sweep structure A, and we produce the intersecting pairs B3:A1 and B3:A2.
In the sixth and last iteration, sweep structure B is empty, so no intersections are produced.

Key Applications

The most common use of the plane sweep algorithm in spatial databases is to perform a spatial join. A spatial join is used to answer questions such as, "list all census blocks within 5 miles of an international airport."

Future Directions

The maximum number of rectangles which the sweep line intersects at any point can be referred to as the maximum overlap. This is an important characteristic for determining the performance of the algorithm. Many GIS datasets have highly skewed data where a large percentage of the data is clustered into small areas. This can greatly increase the maximum overlap.When working with very, very large datasets; the amount of memory available for the storage of the active rectangles can become larger than the available physical memory, which can cause a degradation in performance. In this case, we can either reduce the number of rectangles being compared by integrating spatial indexes (such as R-trees) into our spatial join, or by using more sophisticated adaptations of the plane sweep algorithm.

Cross References

- ▶ Filter and Refine Strategy
- ▶ Minimum Bounding Rectangle

Recommended Reading

1. Baorzsaonyi, S., Kossmann, D., Stocker K.: The Skyline Operator. ICDE (2001)
2. Stojmenovic, I., Miyakawa, M.: An optimal parallel algorithm for solving the Maximal Elements Problem in the Plane. Parallel Computing (1988)
3. Matousek, J.: Computing dominance in E^n. Information processing letters (1991)
4. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest Neighbour Queries. SIGMOD (1995)
5. Hjaltason, G., Samet, H.: Distance Browsing in Spatial Databases. ACM TODS (1999)
6. Kossmann, D., Ramsak, F., Rost, S.: Shooting Stars in the Sky: An Online algorithm for Skyline Queries. In Proceedings of VLDB'02, pp. 275–286 (2002)
7. Papadias, D., Tao, Y., Fu, G., Seeger, B.: An optimal and progressive algorithm for Skyline Queries. ACM SIGMOD (2003)
8. Chomicki, J., Godfrey, P., Gryz, J., Liang D.: Skyline with Presorting. In proceedings of ICDE'03, pp. 717–816, IEEE Computer Society (2003)
9. Lin, X., Yuan, Y., Wang, W., Lu, H.: Stabbing the Sky: Efficient Skyline Computation over Sliding Windows. In Proceedings of ICDE'05, pp. 502–513. IEEE Comp. Soc. (2005)
10. Papadias, D., Tao, Y., Fu, G., Seeger, B.: Progressive Skyline Computation in Database Systems. ACM Trans. Database Syst. 30(1):41–82 (2005)
11. Tan, K.L., Eng, P.K., Ooi, B.C.: Efficient Progressive Skyline Computation. In Proceedings of VLDB'01, pp. 301–310 (2001)
12. Preparata, F.P., Shamos, M.I.: Computational Geometry – An Introduction. Springer, Berlin, Heidelberg, Germany (1985)

Planning Support Systems

► Geocollaboration

Plausibility

► Computing Fitness of Use of Geospatial Datasets

PM-Quadtree

► Quadtree and Octree

Point, Conjecture

► Vague Spatial Data Types

Point, Kernel

► Vague Spatial Data Types

Point Nearest-Neighbor Query

► Nearest Neighbor Query

Point Patterns

► Data Analysis, Spatial

Point Query

► R*-tree

Point, Vague

► Vague Spatial Data Types

Pointless Topology

► Mereotopology

Point-Quadtree

► Quadtree and Octree

Polynomial Spatial Constraint Databases

BART KUIJPERS
Theoretical Computer Science Group, Hasselt University and Transnational University of Limburg, Diepenbeek, Belgium

Definition

The framework of constraint databases provides a rather general model for spatial databases [4]. In the constraint model, a *polynomial spatial constraint database* contains a finite number of relations, that, although conceptually viewed as possibly infinite sets of points in some real space $\mathbf{R}^n$, are represented as a finite union of systems of polynomial equations and inequalities.

Main Text

More specifically, in a *polynomial spatial constraint database*, a relation is defined as a boolean combination (union, intersection, complement) of subsets of some real space $\mathbf{R}^n$ (in applications, typically $n = 2$ or 3) that are definable by polynomial constraints of the form $p(x_1, \ldots, x_n) \geq 0$, where p is a polynomial in the real variables $x_1, \ldots, x_n$ with integer coefficients. For example, the spatial relation consisting of the set of points on the upper half of the unit disk in $\mathbf{R}^2$ can be represented by the formula $x^2 + y^2 < q1 \wedge y \geq 0$. In practice, spatial relations will occur extended with thematic alpha-numeric information, like a name. In mathematical terminology, these spatial relations are known as *semi-algebraic* sets and their properties have been studied extensively [1].

Historical Background

The polynomial constraint database model was introduced by Kanellakis, Kuper, and Revesz [2] in 1990. The application of this model to spatial databases was described by Paredaens, Van den Bussche, Van Gucht [4]. This model was studied extensively in the 1990s and a state of the art book "Constraint databases," edited by G. Kuper, L. Libkin, J. Paredaens appeared in 2000 [3] and the textbook "Introduction to Constraint Databases" by P. Revesz was published in 2002 [5].

Cross References

► Constraint Database Queries
► Constraint Databases and Data Interpolation
► Constraint Databases and Moving Objects
► Constraint Databases, Spatial
► Indexing Spatial Constraint Databases

P

► MLPQ Spatial Constraint Database System
► Visualization of Spatial Constraint Databases

Recommended Reading

1. Bochnak, J., Coste, M., Roy, M.-F.: Géométrie algébrique réelle. Springer-Verlag, Berlin/Heidelberg (1987)
2. Kanellakis, P.C., Kuper, G., Revesz, P.Z.: Constraint query languages. J. Comp. Syst. Sci. **51**, 26–52 (1995)
3. Kuper, G.M., Libkin, L., Paredaens, J. (eds.): Constraint Databases. Springer-Verlag, Berlin/Heidelberg (2000)
4. Paredaens, J., Van den Bussche, J., Van Gucht, D.: Towards a theory of spatial database queries. In: Proceedings of the 13th ACM Symposium on Principles of Database Systems, pp. 279–288 (1994)
5. Revesz, R.Z.: Introduction to Constraint Databases. Springer-Verlag, New York (2002)

Polynomials

► Biomedical Data Mining, Spatial

Polynomials, Orthogonal

► Biomedical Data Mining, Spatial

Populating, Topology

► Oracle Spatial, Geometries

Population Distribution During the Day

BUDHENDRA BHADURI[1]
Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Synonyms

Daytime population; Mobile population; Nonresidential population

Definition

Population distribution during the day can be defined as the distribution of population in an area during the daytime hours. However, a precise definition of daytime hours is challenging, given the geographic variability in the length of a day or daylight hours. The US Census Bureau used "normal business hours" as the span of time to describe daytime population [1]. Given that censuses typically estimate residential population, it represents a nighttime population distribution. In that respect, the daytime population in an area may be broadly defined as distribution of population at times other than when they are expected to be at their residences at night which extends the duration from business hours to include the evening hours as well.

Historical Background

Population data has served as a fundamental backbone for planning sustainable development. There is evidence from the early 1600s of a population census in Virginia where people were counted in nearly all of the British colonies that became the United States at the time of the Revolutionary War [2]. Historically, it has been used to meet a variety of long term socioeconomic and political planning needs. For example, the first US Census of 1790, which counted 3.9 million residents, helped raise the membership in the US House of Representatives from an original 65 to 105. Modern censuses that include not only population count, but also its demographic and socioeconomic characteristics have had a tremendous impact on many aspects of our society covering, among other things, urban planning and housing development, transportation planning, energy demand and infrastructure planning, healthcare planning, environmental impact assessment, emergency preparedness and response, and scientific research. However, the majority of these planning activities have been aimed at medium- to long-term solutions over a number of years and consequently a general geographic assessment of population, described through their residential locations, was adequate to address such planning processes. Movement of population during a day results directly from people traveling to the locations of their daytime activities (employment, business, educational institutions, and recreational locations), away from their residences [3]. The patterns of such population displacements depend on the relative geographic distribution of residential and business areas. In most modern societies, these two activity locations are distinctly separated in space, and employment or business locations contain fewer residences than businesses. Consequently, a large number of people move into these areas while only a few leave, resulting in a substantial swelling in the daytime population of that area. The motivation to formalize the concept of nonresidential and daytime population distribution is rooted predominantly in two areas. First, it is widely understood that analysis of the daytime population distribution provides a very competitive economic advantage, as business-

[1]The submitted manuscript has been authored by a contractor of the U.S. Government under contract DE-AC05-96OR22464. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

es are enabled to target specific consumer bases depending on their locations and convenience of access during that majority of the 24-h period when people are out of their residences. In recent years, a stronger requirement for understanding daytime population has emerged from the emergency preparedness and response community to assess the at-risk population from the threats of technological and natural disasters and deliberate attacks on human lives such as terrorist events.

Scientific Fundamentals

Population movement is a function of both geographic space and time. The mobility of a population during the day is driven by people's need to temporarily relocate for activities such as education (schools, colleges, universities), employment, businesses (shopping, post offices, restaurants, and others), or recreation (parks, museums, other tourist attractions). In general, the daytime population distribution of an area can be conceptually expressed as:

$$\begin{aligned}\text{Daytime Population} = {} & \text{Workers} + \text{School children} \\ & + \text{Tourists} + \text{Business travelers} \\ & + \text{Residual Nighttime Residential Population}\end{aligned}$$

or,

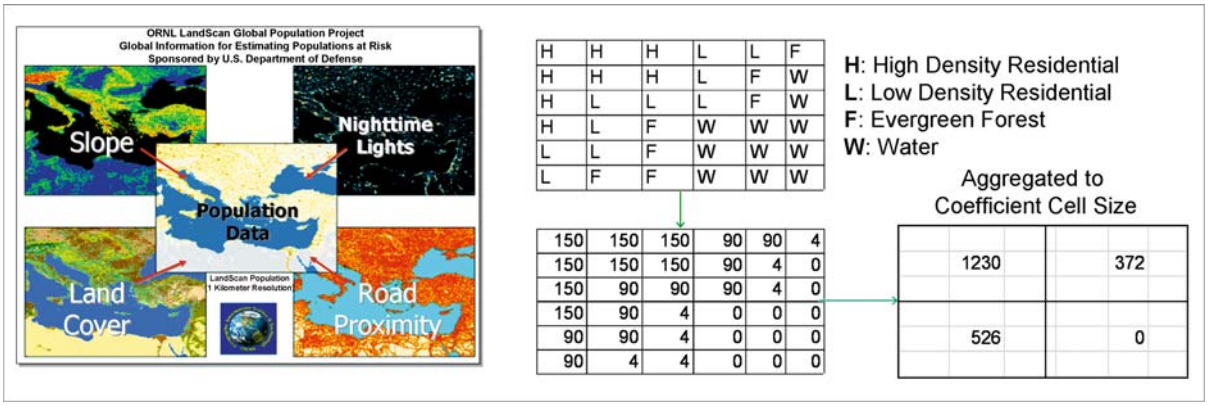
$$\begin{aligned}\text{Daytime Population} = {} & \text{Nighttime Population} \\ & + \text{Daytime incoming population} \\ & - \text{Daytime outgoing population}\end{aligned}$$

Deriving a quantitative estimate from the above qualitative expressions involves further analyses of population data, which can be represented as:

$$\begin{aligned}\text{Daytime Population} = {} & \\ & \text{Nighttime Residential Population} \\ & - \text{Workers leaving during the day} \\ & + \text{Workers moving in during the day} \\ & - \text{School children leaving during the day} \\ & + \text{School children moving induring the day} \\ & + \text{Tourists visiting during the day} \\ & + \text{Businesstravelers coming into the area}\end{aligned}$$

Although this may not be the most accurate representation, it is largely accepted as the expression leading to the best available daytime population estimates. It follows that the development of quantitative estimates of daytime population distribution involves two distinct components. The first component involves identification of daytime activity locations such as businesses, schools, and other recreational activities. The second component covers the identification and distribution of the mobile population that are at those locations. Usually, it is easier to gather information on the first as these are static geographic features and are commonly captured in public and commercial databases for various infrastructures, or can be derived from remote-sensing-based land-cover data, high-resolution satellite and aerial photographs, or state and local government data. However, it is extremely challenging to obtain information on the number and nature of movement of people during the day that comprehensively captures the net displacement of the nighttime residential population during daytime. Although, detailed population movement data sets may be available for isolated local communities, they are not available at a national scale. In fact, the US Census Bureau's compilation of journey-to-work data is the only readily available and nationally consistent data set for the US that describes people's movement from residences to employment locations. Consequently, the US Census Bureau's estimate of daytime population based on the 2000 Census only reflects populations based on travel to work. Similarly, it does not limit the work-related commuting to specific hours. All worker-related travel, irrespective of what time of the day it occurs, has been used to derive these estimates of daytime population [1].

An important aspect of daytime population distribution is the geospatial scale at which it is estimated. Theoretically, the finest spatial resolution achievable through the map algebra technique described above is directly tied to the finest scale of the available input data. For example, the US Census Bureau collects worker commuting data at the census tract level and reports national daytime population distribution at the county level. It also reports estimates of daytime population for key cities in each state. Similar city level estimates of daytime population from government and commercial sources are available for Japan [4], Canada, and the US. All these data sets appear to be heavily focused on worker population movement during the day and the data is presented through vector data models (points and polygons). For example, daytime population fluxes are restricted to individual county and city boundary polygons. Some commercial databases represent individual activity locations as points that potentially offer high spatial accuracy but mostly account for worker population at individual business locations. In reality, the data sets necessary to comprehensively estimate daytime population exist in the forms of points and polygons, which makes it challenging to create a high-resolution population distribution through simple map algebra analysis. It requires integration of disparate spatial data and advanced geospatial modeling where the spatial mod-



Population Distribution During the Day, Figure 1 Example illustrating the use of land-cover data in the LandScan dasymetric model

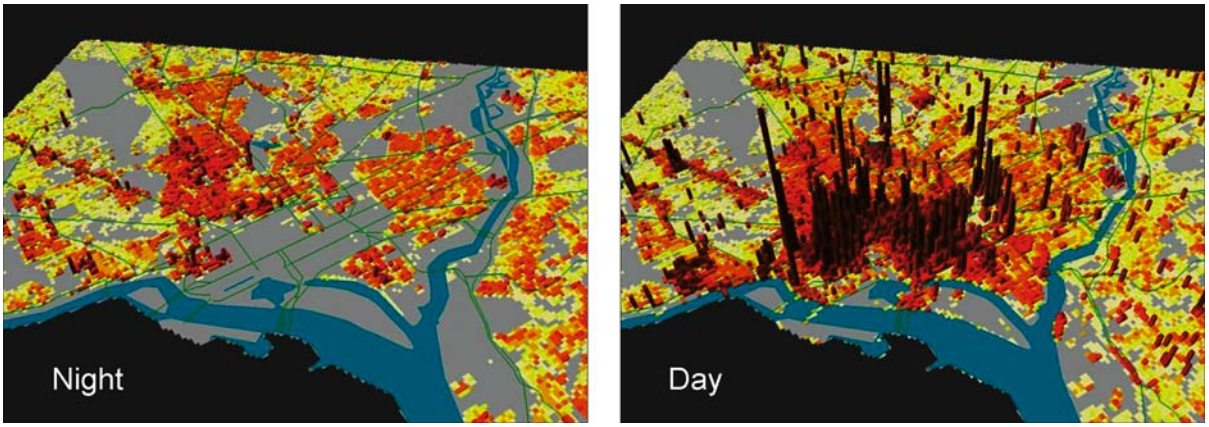
el enables decomposition of the input data into finer spatial resolutions and then representation through uniform raster or gridded datasets.

Decomposition of population distribution estimates has been a well-known problem. Several interpolation and decomposition methods have been developed to address this issue with census (polygonal) population data. They include areal weighting, pycnophylactic interpolation, dasymetric mapping, and various smart interpolation techniques. Areally weighted interpolation is the simplest of the methods, where a regular grid is intersected with the Census polygon and each grid cell is assigned a value based on the proportion of the polygon contained in each cell [5,6,7]. This method implies an assumption of uniform distribution of population, which is not a realistic solution for decomposition of population data. Pycnophylactic interpolation extends the areal weighting methodology by iteratively applying a smoothing function to the raster cell values, with the weighted average of its nearest neighbors, while preserving the total population count of the polygon [8]. This method creates a continuous surface which contradicts the obvious discontinuous nature of population distribution. Dasymetric modeling is analogous to areal interpolation but uses ancillary spatial data to aid in the interpolation process. The ancillary spatial data is at a finer spatial resolution and the variability in its values enables an asymmetric allocation of population values. Land cover/land use is the best example in this respect, where different land cover or land use categories for each cell can be used as weighting functions for population distribution such that urban areas will have a higher weight than forested areas (Fig. 1) [7,9,10]. Smart interpolation, in principle, is a multidimensional version of a dasymetric model where the allocation refinement comes from more than one ancillary data sources, which are at a finer resolution than the population polygon [10,11]. The utility of such interpolation techniques at local scales are well documented. In fact, there are two well-known publicly available data sets, the Gridded Population of the World (GPW) [12] and the LandScan Global Population database [13], which employ such a method to produce global population distribution data. While GPW is a product of simple areal weighting interpolation at 2.5 arc-min or approximately 5-km cell size, LandScan, at 30 arc-s or approximately 1-km cell size is the finest resolution global population data available to date, derived through advanced spatial data integration and multidimensional dasymetric modeling or smart interpolation.

Although both GPW and LandScan datasets are developed using Census information, GPW depicts a nighttime residential population (i. e., directly decomposed Census data) while the LandScan database represents an "ambient" or average of the 24-h period population (because the model assigns some parts of the populations based on nonresidential activities). Both databases are publicly available for noncommercial usage from Columbia University [14] and Oak Ridge National Laboratory [15] respectively. GPW is updated periodically, while LandScan Global Population database has been updated and released annually since 2000.

The development of daytime population distribution models and databases is significantly more challenging, as it requires further integration and modeling of activity-based datasets into the residential population distribution model. In 2004, the US Census Bureau released the following three daytime population distribution data tables based on the 2000 census [1]:

- Table 1. Leading Places on Percent Change in Daytime Population, by Size (202 highly populated cities)
- Table 2. The United States, States, Counties, Puerto Rico and Municipalities
- Table 3. Selected Places by State (6524 communities)



Population Distribution During the Day, Figure 2 Difference in population distribution between nighttime and daytime as illustrated by LandScan USA data for Washington DC

However, these data sets only take into account the commuting worker population in an area. The best spatial resolution of these data is still at the community level (small cities) and thus is appropriate for general-purpose planning. Expanding on their LandScan Global Population research program, since the early 2000s, the US Department of Energy's Oak Ridge National Laboratory has played a pioneering role in developing an advanced scalable daytime population distribution model for the nation called LandScan USA [16,17]. At an unprecedented resolution of 3 arc-second or approximately 90-m cell size, LandScan USA demonstrates a consistent methodology for developing daytime population distribution (Fig. 2). This enhancement in resolution comes from the incorporation of a large number of high-resolution ancillary data sets used in the LandScan USA dasymetric model. Some of these ancillary data sets include:

- Population
  - Census block population
  - Census tract-to-track worker flow
  - Bureau of Labor Statistics quarterly updates for worker population.
- Roads
  - Tele Atlas North America (formerly known as Geographic Data Technology) Dynamap
  - US Census TIGER data
  - Navigational Technologies (NAVTEQ) roads
- Land Cover/Land Use
  - National Land Cover Data (NLCD)
  - State and local level GIS databases
- Slope
  - National Elevation Data (NED)
- Academic Institutions
  - Department of Education
  - Environmental Systems Research Institute(ESRI)
  - Tele Atlas North America;
- Prisons
  - Department of Justice National Jail Census
- Hospitals
  - American Hospital Association (AHA)
- Business Employment
  - ESRI Business Database (Info USA)
- Ortho Imagery
  - Google Earth
  - Earth Viewer
  - Microsoft Terra Server

In addition to the worker population in an area, LandScan USA database accounts for children of K–12 school age, university students, institutional population (jails and prisons), and a mobile daytime residential population at various activity locations such as shopping malls, post offices, cultural attractions, and recreational facilities (parks). Development of LandScan USA version 1.0 has been completed for the 50 US States and Puerto Rico but the data is not publicly available yet. It should be noted that both LandScan Global and LandScan USA are evolving databases and new ancillary input data sets are continuously added to the model as they become available.

A similar approach [18] has also been adopted for estimating daytime population at a lower resolution of 250-m grid cells. The coarser resolution of this data set has been attributed to coarser resolution of input variables in the model such as county-to-county worker mobility data from Census as compared to tract-to-tract worker mobility data used in LandScan USA. Moreover, this estimation is solely based on worker and residential populations and does not account for populations at academic institutions, commercial retail locations, and recreational areas.

Key Applications

Missions of national priority ranging from socio-environmental studies to homeland security utilize population data as one of the critical elements. High-resolution population distribution data during daytime hours is even more significant for successfully addressing research and practical applications that require an estimation of mobile population. Applications of daytime population distribution are numerous and can be broadly divided into two categories.

Estimating Population at Risk from Disasters

Large numbers of human lives are at risk from natural and technological disasters. Volcanic eruptions, earthquakes, hurricanes, floods, wildfires, blizzards, droughts, and tornadoes are examples of natural disasters that are slow in their onset, predictable in most cases (except earthquakes and volcanic eruptions), and typically geographically restricted. However, natural disasters are unpreventable and for the most part uncontrollable. Technological disasters, on the other hand, occur suddenly (i. e., are unpredictable) but can be controlled and their impact minimized through effective disaster planning and management. Examples of technological disasters include explosions, electrical blackouts, nuclear accidents, and bioterrorism. Critical application domains for estimating the population at risk include national and homeland security, where improved knowledge of where people live relative to sites of potential terrorist activities can refine estimates of potential populations exposed [18] or injured for rapid risk assessment; and emergency preparedness and response where the daytime population can support emergency response resource planning, emergency evacuation planning, and disaster relief delivery.

Public Health and Socioeconomic Analysis

Public health is probably one of the most promising areas that can take great advantage of daytime population distribution information. Disease epidemiology with short- and long-term exposure assessment and evaluating access to health care facilities and locating future health care facilities can be done effectively with an understanding of daytime population distribution. Assessment of the mobility daytime population with respect to their residences also facilitates understanding of contagious disease propagation patterns. Exposure and risk assessment from environmental pollutants for work-related activities can be performed very effectively using the daytime distribution of population. Given that workers are likely to spend 50% or less of their time at home during work days, the daytime population distribution provides a tremendous advantage for occupational exposure analysis over using traditional census data. High-resolution daytime population data also reduces population distribution errors around point and area sources and can be of significant help for environmental justice analysis. Other socioeconomic applications include demographic analysis to evaluate socioeconomic disparity patterns of a region in terms of work-related commuting patterns for different demographic groups and estimating rates and trends of urban sprawl.

Future Directions

Accurate estimation and representation of daytime population distribution poses significant challenges. First, any region witnesses an influx of tourists (or visitors) and business travelers during the day, particularly in large urban areas and cultural/natural attractions (such as national parks). In addition, a large number of people travel along roads driving through major urban areas. Such transitional population is not effectively captured in any consistent and organized databases and will require advanced data integration and modeling techniques to be effectively included in a daytime population distribution. Current spatial modeling and population distribution techniques only locate population at specific activity locations and do not account for the commuting time when the mobile population is on the transportation networks. Thus, a true average daytime population distribution requires details of worker commuting patterns and non-worker travel habits, which includes data for the number of people at necessary service locations such as post offices, banks, shops, and parks. Another important aspect of daytime population distribution is to characterize the temporal variability. The nature of daytime population distribution can be significantly different depending upon whether the data represents a working or a weekend day or holiday. Moreover, there is also a seasonal and weather impact on the daytime population distribution. Summer days do not have students in academic institutions and have more people at outdoor locations. In contrast, a larger population tends to be indoors during days with weather extremes (such as extreme heat, cold, or storms). Thus assessment of a true "representative" daytime population of a region will require development of an average distribution from such different daytime population distribution scenarios.

Cross References

- ► Computing Fitness of Use of Geospatial Datasets
- ► Data Analysis, Spatial
- ► Geodemographic Segmentation
- ► Geographic Dynamics, Visualization And Modeling
- ► Homeland Security and Spatial Data Mining

► Intelligence, Geospatial
► Movement Patterns in Spatio-temporal Data

Recommended Reading

1. US Census Bureau: Census 2000 PHC-T-40. Estimated Daytime Population and Employment-Residence Ratios: 2000; Technical Notes on the Estimated Daytime Population. http://www.census.gov/population/www/socdemo/daytime/daytimepoptechnotes.html (2000). Accessed 17 Sept 2007
2. US Census Bureau: US Census Bureau's Fact Finder for the Nation: History and Organization. http://www.census.gov/prod/2000pubs/cff-4.pdf (2000). Accessed 17 Sept 2007
3. Quinn, J.: The daytime population of the central business district of Chicago. Review by Breese, G.W. Am. Sociol. Rev.**15**, 827–828 (1950)
4. Japanese Statistics Bureau, Ministry of Internal Affairs and Communications: Population Census, Daytime Population http://www.stat.go.jp/english/data/kokusei/2000/jutsu1/00/01.htm (2000). Accessed 17 Sept 2007
5. Goodchild, M., Anselin, L., Deichmann, U.: A Framework for the Aerial Interpolation of Socioeconomic Data. Environ. Planning A **25**, 383–397 (1993)
6. Goodchild, M., Lam, N.: Aerial interpolation: a variant of the traditional spatial problem. Geo-Processing **1**, 297–312 (1980)
7. Mennis, J.: Generating surface models of population using dasymetric mapping. Prof. Geogr. **55**, 31–42 (2003)
8. Tobler, W.: Smooth pycnophylactic interpolation for geographical regions. J. Am. Stat. Assoc. **74**, 519–530 (1979)
9. Wright, J.: A method of mapping densities of population: with Cape Cod as an example. Geogr. Rev. **26**, 103–110 (1936)
10. Langford, M., Unwin, D.: Generating and mapping population density surfaces within a geographical information system. Cartogr J. **31**, 21–26 (1994)
11. Cohen J., Small, C.: Hypsographic demography: the distribution of human population by altitude. Proc. Natl. Acad. Sci. **95**, 14009–14014 (1998)
12. Deichmann, U., Balk, D., Yetman, G. Transforming Population Data for Interdisciplinary Usages: from Census to Grid. http://sedac.ciesin.columbia.edu/gpw-v2/GPWdocumentation.pdf (2001). Accessed 17 Sept 2007
13. Dobson, J., Bright, E., Coleman, P., Durfee, R., Worley, B.: LandScan: a global population database for estimating populations at risk. Photogramm. Eng. Remote Sensing **66**, 849–857 (2000)
14. Center for International Earth Science Information Network: Gridded Population of the World (GPWv3). Columbia University, New York. http://sedac.ciesin.org/gpw/ (2007). Accessed 17 Sept 2007
15. Oak Ridge National Laboratory: LandScan Global Population Project. Oak Ridge National Laboratory, Tennessee. http://www.ornl.gov/sci/landscan/ (2007). Accessed 17 Sept 2007
16. Bhaduri, B., Bright, E., Coleman, P., Dobson, J.: LandScan: locating people is what matters. Geoinformatics **5**, 34–37 (2002)
17. Cai, Q., Rushton, G., Bhaduri, B., Bright, E., Coleman, P.: Estimating small-area populations by age and sex using spatial interpolation and statistical inference methods. Trans. GIS **10**, 577–598 (2006)
18. McPherson, T., Brown, M.: Estimating daytime and nighttime population distributions in U.S. cities for emergency response activities. In: Proceedings of the Symposium on Planning, Nowcasting, and Forecasting in the Urban Zone, 84th AMS Annual Meeting, Seattle WA, 11–15 Jan 2004
19. Dobson, J., Bright, E., Coleman, P., Bhaduri, B.: LandScan2000: A New Global Population Geography, Remotely-Sensed Cities. Taylor and Francis, London, UK (2003)

Position, Absolute

► Indoor Localization

Position Location

► Indoor Positioning with WirelessLocal Area Networks (WLAN)

Position, Relative

► Indoor Localization

Positional Accuracy Improvement (PAI)

CARSTEN RÖNSDORF
Ordnance Survey, Southampton, UK

Synonyms

PAI; Synchronization of spatial data; Map quality; Map overhaul; Digitization of maps; Relative positional accuracy; Absolute positional accuracy; Rubber sheeting; Geometric fidelity; Shifting geometries; Conflation

Definition

Positional Accuracy Improvement is the process of improving the position of the coordinates defining the geometry of a feature in a geospatial dataset to better reflect its "true" position. This "true" position can either relate to the absolute position in an overarching Coordinate Reference System such as WGS-84 or ETRS-89 or to the relative position in relation to the geometry of another feature in the vicinity.

The PAI process is commonly utilized in two different, but related ways:

PAI of Reference Data

This process deals with improving the position of geometries in a reference dataset that describes physical or abstract features of the earth. These reference datasets are typically large-scale cadastral or topographic datasets issued by National Mapping Organizations but can

also consist of other regional or local datasets, usually described as "land bases" and used to reference other information.

PAI of User Data

If a geospatial vector dataset is derived from a reference dataset either by digitizing or copying geometries, both datasets have a topological relationship in the sense that the former is based on the latter. In regards to user datasets PAI describes the subsequent synchronization of one or more of the user datasets with the already positionally improved reference dataset in order to re-instate the relationships between geometries.

Historical Background

Mapping as a technology and means of communication about entities of the real world has a history that can be traced back to at least 3500 BC [1]. The quality of a map is directly related to the methods of surveying and cartography used at the time when the map was produced. It is common in surveying and mapping to avoid a complete recollection of the data every time a new map of a previously surveyed area is created. It is good practice to use old maps as a basis and update and enhance them with new and additional information. This technique, which can be described as "map overhaul", is used to create new themes as well as updating a map to a more current status.

Many geospatial datasets that are used in today's digital environments are updated in the same way. In fact, many of these datasets were originally digitized from paper maps and subsequently updated. This means that many geospatial datasets in use today are an amalgamation of data from different sources, integrated at different times by different methodologies.

With the move from paper maps to geospatial datasets the ability to easily combine spatial data from different sources has dramatically increased. This allows various independently created datasets to be jointly presented and analyzed. The spatial integration of data from various sources requires an understanding about the positional accuracies of the geometries in the datasets to avoid mismatches and misinterpretations. Positional accuracy is an important quality aspect of geospatial datasets, a viewpoint that is underpinned in the data quality description in ISO19113 [2].

To give an example: A dataset digitized from a small-scale map will have a radically different accuracy than another one surveyed by differential GPS. Geospatial reference datasets, particularly large-scale topographic and cadastral data issued by Local or National Mapping Organizations, were typically created over the course of decades or centuries. Varying accuracies have their origin in different surveying and transformation methods that were applied to the data after their initial creation.

Positional Accuracy Improvement as a term was initially introduced by Ordnance Survey of Great Britain to describe its coordinated program to improve the positional accuracy of large scale topographic data [3]. This program was one of the major activities to obtain a large-scale reference dataset to express the topography of the whole of Great Britain and is used as a case study later in this chapter.

A series of subsequent publications and workshops [4,5] have established the term PAI to articulate systematic changes to improve the position of geospatial reference data and its effect on datasets that were derived from these. It was found that the issues that lead to PAI may vary in different national contexts, but that the core technical problem is often identical [4].

Scientific Fundamentals

The positional accuracy of geodata can be described by the terms relative and absolute accuracy, which are defined as follows.

Relative Positional Accuracy, which has traditionally been used to indicate the positional accuracy of maps, is defined as

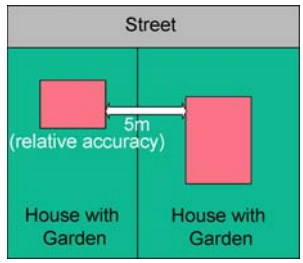
> the difference of the distance between two defined points in a geospatial dataset and the true distance between these points within the overall reference system.

Practically, the true distance can be measured using conventional terrestrial surveying techniques, such as a tape or laser distance measure, and can be compared to the calculated length between the two data points. An example for relative positional accuracy is given in Fig. 1.

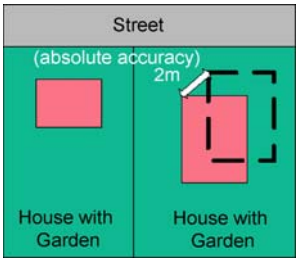
In the 1980s satellite navigation technology, such as GPS, introduced the possibility of obtaining a point's coordinate directly without relating to neighboring features. Therefore another accuracy definition is needed.

Absolute Positional Accuracy is defined as

> the distance between a defined point in a geospatial dataset and its true position in the overall reference system.



Positional Accuracy Improvement (PAI), Figure 1 Relative positional accuracy



Positional Accuracy Improvement (PAI), Figure 2 Absolute positional accuracy

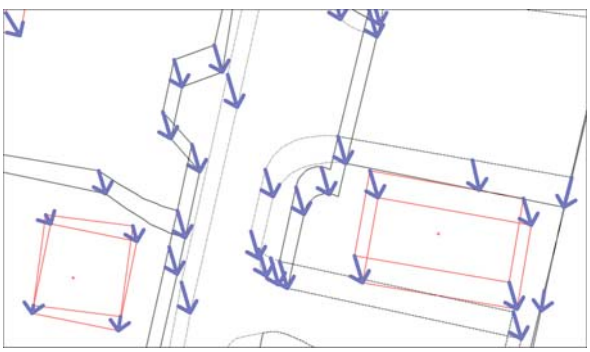
Practically, the true position within the reference system can be determined to centimeter accuracy using differential GPS surveying. An example for absolute positional accuracy is given in Fig. 2.

Both the absolute and relative positional accuracy of a given dataset can be determined by calculating the differences between the mentioned distances of a significant sample of the dataset and can be expressed as a root mean square error (RMSE) that relates to a one sigma (standard deviation) confidence level. Assuming a Gaussian distribution for the errors or differences, this means that the likelihood that a measurement falls into the range expressed by the RMSE—plus or minus one meter, for example—is 67%.

Historically, relative accuracy has been more important to most users than absolute accuracy since it expresses a local quality statement of a geospatial dataset taking the relationship of neighboring features into account.

Topological Relationships Between Geometries

Vectors between identical points of geometries in an unimproved and an improved (or shifted) datasets are called **Link Vectors**. The start and end-points of these vectors have coordinate values in the overall reference system. This means that they describe the shift between the points in both datasets. A **Link Vector Field** is a collection of link vectors for a dataset or part of a dataset. An example for a link vector field connecting identical points in two displaced sets of geometries is displayed in Fig. 3.



Positional Accuracy Improvement (PAI), Figure 3 Link vector field

The process of altering the coordinate values of points in a datasets by adding link vectors to individual coordinates is called **Rubber Sheeting**. If the link vector field does not contain a link vector that originates on the point to be shifted, a link vector is interpolated within the link vector field. Commonly used interpolation algorithms include nearest neighbor, inverse distance weighting or natural neighbor [6].

The term **Geometric Fidelity** is used to describe how well the shape of a line or area geometry is retained in a PAI process [7,8]. Geometric fidelity is defined as

> the difference between the shape of a geometry or set of geometries as an ordered sequence of points in a geospatial dataset and the sequence of the corresponding points in the real world.

Practically, the geometric fidelity can be assessed for each geometry either by visually checking the distortion of a geometry or by calculating and comparing the angle between edges in the dataset and the real world. The latter can be approximated by GPS measurements of the end points of the lines.

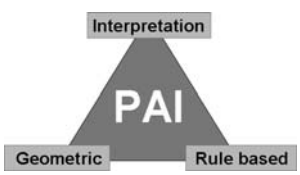
The process of rubber sheeting will decrease the geometric fidelity of a dataset in a PAI process in most cases. The geometric fidelity will not be altered if the link vector field leads to a close approximation of a linear shift and or a rotation.

Shifting Geometries

In addition to re-surveying or re-capturing data, there are three fundamental ways to move data in a PAI scenario. They are illustrated in Fig. 4, the **Shifting Triangle**.

Shifting by Interpretation describes the method of moving features to their improved position by determining this position on an individual basis for every geometry. It is based on the perceived relationship between these features and others in the vicinity. The new position is determined by a human operator or, alternatively, an artificial intelligence process, by interpreting the feature according to an intangible or tangible capture specification.

Geometric shifting relates to the application of a link vector field that describes the difference between the old and the improved base data by means of geometric transformations. Rubber sheeting is a commonly used example for geometric shifts.



Positional Accuracy Improvement (PAI), Figure 4 Shifting triangle

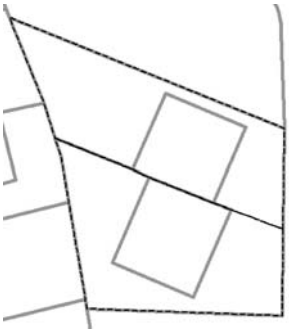
Rule based shifting make use of the specification of the dataset and explicit relationships between features in the base and user datasets. These can be expressed by geometric constraints, such as preserving right angles between lines or calculating distance functions between user and reference geometries [9]. The distance function is described in a later paragraph.

The shifting triangle implies that the most efficient process to migrate user data is likely to be a combination of the three fundamental methods. An often used combination would be to rubber sheet the data first, then 'snap' relevant features back to the improved reference dataset according to a rule, followed by an interpretational correction of geometries that were incorrectly shifted.

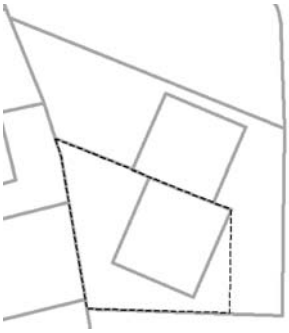
Topological Relationships Between Geometries

The following three figures illustrate important relationships between reference data (thick grey lines) and user data (thin dashed lines) using polygon geometries as an example.

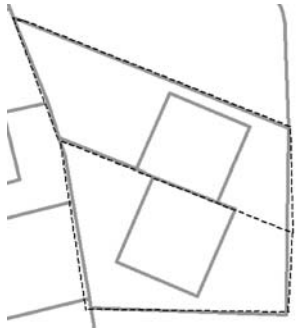
In Fig. 5 the user data completely follow the geometry of the reference data – the vertices of the user data are identical to the vertices of the reference data. Within a Geographic Information System many of this data are created by digitization using a 'snapping' algorithm. Hence these data can be described as 'snapped' user data. In Fig. 6 the user data contain geometry that is not in the reference. Therefore this part of the polygon can't be linked to any geometry in the reference data. Figure 7 depicts a common scenario where the user data are digitized against the reference data without using a 'snapping' algorithm. In this case the user data vertices are not identical with the reference data vertices, but close to them. Since the quality of digitization could be better (as illustrated in Fig. 5), this relationship can be described as roughly digitized user data.



Positional Accuracy Improvement (PAI), Figure 5 Snapped user data



Positional Accuracy Improvement (PAI), Figure 6 Partly snapped user data

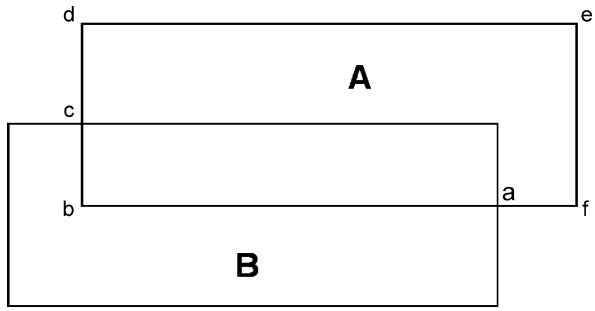


Positional Accuracy Improvement (PAI), Figure 7 Roughly digitized user data

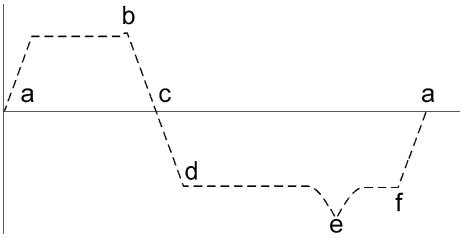
Distance Function

A spatial relationship between two polygons can be described by a distance function [10]. The distance function reports the minimum distance separating two polygons as a function $f(n)$ of the perimeter n of polygon A. Where two polygons are coincident, the function will report a zero minimum distance. $f(x) > 0$ indicates that the perimeter of polygon A falls within polygon B and therefore overlaps while $f(x) < 0$ shows that the perimeter of polygon A falls outside polygon B.

For the two polygons A and B, as shown in Fig. 8, the distance function between the boundary of polygon A and polygon B is displayed in Fig. 9. It is created by following the perimeter of polygon A from intersection point *a* over *b, c, d, e* and *f* back to point *a*. For all points on the perimeter (or a number of points that are placed in small discrete intervals on the perimeter) the shortest distance to polygon B is calculated.



Positional Accuracy Improvement (PAI), Figure 8 Two simple polygons



Positional Accuracy Improvement (PAI), Figure 9 Distance function expressing the relationship between the polygons in Fig. 8

The distance function for snapped user data (see Fig. 8) constantly equals zero, while the distance function for partly snapped polygons (see Fig. 9) will be zero for the snapped part of the polygon and have a distinctive peak where user data does not follow reference geometries. For the scenario shown in Fig. 7, the roughly digitized data, the distance function will be close to zero, with some noise indicating the difference between the user data and underlying reference data [9].

A term related to PAI is **Conflation**, meaning automated map compilation [11]. This technique, pioneered in 1985, allows the integration of two overlapping geospatial datasets by matching some corresponding structures in both sets and subsequent transformation of all features into one dataset. While conflation is a technique that can be utilized to execute PAI, PAI itself focuses on maintaining the synchronicity between two or more datasets following geometry changes to the reference dataset and allows a number of different technical solutions.

A fundamental discussion of the Mathematical Models for Geometrical Integration that are the basis for PAI as well as links to additional literature can be found in an article by Kampshoff [12].

Key Applications

Positional Accuracy Improvement introduces geometry changes to geospatial datasets in order to improve their relative and/or absolute positional accuracy.

While an improvement of the relative accuracy is a possible scenario, today's main application area is the improvement of the absolute accuracy within a national or global coordinate system. The influence of new surveying methods, namely the impact of Global Navigation Satellite Systems, on the production of large-scale topographic and cadastral datasets has been the main driver to improve the positional accuracy of these datasets. New surveys are often very accurate within a global Coordinate Reference System and need to be integrated into existing topographic or cadastral reference datasets. A large part of the existing data is typically based on historic surveys in local reference systems of a lesser accuracy.

Positional Accuracy Improvement (PAI), Figure 10 Reference and user data

Reference Data and User Data

In practical terms geospatial data can usually be divided into two categories: reference data that provide the geospatial context and user data that comprise additional features supporting a particular application. In many cases the user data are generated by the users themselves while the reference data are provided by a National Mapping Agency or private data provider. On a paper map both the base map and the user data are usually drawn or printed onto the same surface and cannot easily be separated from each other. Digital datasets maintain both as two or more completely separate layers that are just combined for analysis and publication on the screen or in print. An example for reference and user data is given in Fig. 10. The thick, grey lines represent reference data; the thin, red lines indicate user data.

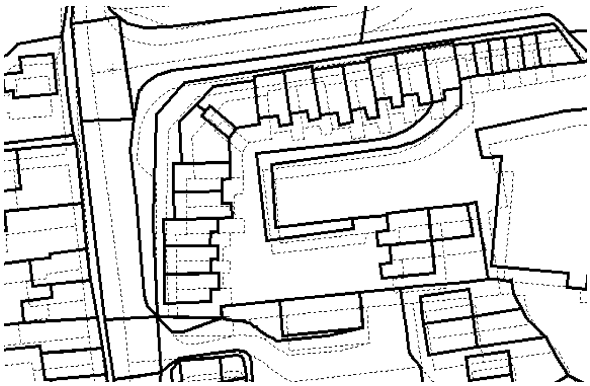
PAI on Reference Data

The improvement of an existing reference dataset can be done by re-surveying the geometries, shifting them or a combination of both. Figure 11 gives an example for the difference between an unimproved (thick line) and an improved (thin, dashed line) reference data set.

After a reference dataset is shifted, all future updates of that dataset will typically apply to the improved dataset. This usually triggers a PAI process on user data in conjunction to datasets that were derived from this reference dataset.

PAI on User Data

User datasets that were derived from or created to be used in conjunction with a particular reference dataset may not



Positional Accuracy Improvement (PAI), Figure 11 Reference data before (*bold line*) and after applying PAI (*dashed line*)

be synchronized with that reference dataset after the latter one has been improved. This means that the topological relationships between geometries in the reference dataset and the user data may be destroyed.

This can have a significant impact on the use of these datasets. Applications such as automated searches for conveyancing, may produce incorrect results if a search is done using improved reference data in conjunction with unimproved user data.

Once a user dataset is shifted to its post-PAI position, geometric interoperability between this dataset and the improved reference is re-established and both datasets can be used in conjunction again.

Use Case: PAI in Great Britain

In Great Britain, the original topographic surveys date as far back as the early 1800s. More importantly, a large amount of the surveys that form the backbone of today's large-scale digital reference data was acquired during the first half of the 20th century. At that time it was common practice to use separate, county-specific reference and coordinate systems to survey and display the maps (known as County Series maps). A fundamental approach to integrate those projections into one common metric coordinate system for Great Britain, the British National Grid, was started around 1938 and finished after the Second World War. Today all large-scale reference data are held in a national geographic database that currently holds about 440 million features as well as selected spatial relations between features.

It was apparent, even before GPS was used as a surveying tool, that new topographic details could not always be seamlessly integrated into the national geographic database, resulting in operational overheads to maintain the database as well as using the data in conjunction with user data.

While differential GPS methods deliver absolute positional accuracies of 10 centimeters or better, features in large-scale Ordnance Survey data have an absolute positional accuracy of between 2.8 m RMSE in rural areas and 0.4 m RMSE in urban areas. This indicates the accuracy of the absolute position of a coordinate in the context of the British National Grid coordinate system. In contrast to this, the relative positional accuracy has always been significantly better.

Following earlier debates that go back to the 1970s, Ordnance Survey started to plan a national program to improve the absolute positional accuracy of its rural large-scale data in the late 1990s. It applies to 152,000 km^2 (or about two thirds of the area of Great Britain) and excludes the major urban areas, which were already resurveyed to a higher standard from 1947 onwards, as well as mountain and moorland regions, where improving the positional accuracy is not necessary or economically viable. The program was completed in March 2006 and is future-proofing the value of the national geographic database. The absolute positional accuracy of the data after the improvement is 1.1 m RMSE in rural areas and 0.4 m RMSE in selected rural towns.

In Great Britain this data set is widely used as a reference in conjunction with individual user datasets by several hundred organizations throughout the country. The need to improve the positional accuracy of associated user datasets has created at least 30 different solutions and services to shift user datasets [13].

Use Case: MAF/TIGER Accuracy Improvement in the United States

In the United States the most prominent PAI program is undertaken by the US Census Bureau as part of the MAF/TIGER Accuracy Improvement Project [14]. The program aims at improving the accuracy of the TIGER (Topologically Integrated Geographic Encoding and Referencing System) database to 3.8 meters RMSE for all 50 states as well as Puerto Rico and the U.S Virgin Islands. Unimproved data have been reported to differ up to 150 meters from its (true) Differential-GPS position. This will allow the Bureau to match geographic locations to census geographies in a more automated way.

Future Directions

It is anticipated that the widespread use of GPS and aerial photography that is rectified against GPS control points will significantly increase the importance of absolute accuracy. Therefore the need to positionally improve datasets with a low positional accuracy is likely to increase in order

to make them geometrically interoperable with datasets of a higher absolute positional accuracy.

To date most PAI programs were implemented as coordinated programs. These programs usually focused on the improvement of a particular dataset and were focused on the individual requirements for geometric interoperability of this dataset with those of a higher positional accuracy. In the future the development of a more generic process to improve the accuracy of geospatial datasets would be beneficial. Particularly the implementation of PAI as a web service, performed on the fly, as the data are being prepared or loaded into a client application, will be a powerful tool in order to make datasets geometrically interoperable.

Cross References

This article has been prepared for information purposes only. It is not designed to constitute definitive advice on the topics covered and any reliance placed on the contents of this article is at the sole risk of the reader.

► Data Infrastructure, Spatial

Recommended Reading

1. Dorling, D., Fairbairn D.: Mapping: ways of representing the world. Longman, London (1997)
2. International Organization for Standardization: ISO 19113 Geographic information – Quality principles. Geneva (2002)
3. Havercroft. M.: Consultation paper: Positional accuracy of large-scale data and products. Ordnance Survey, Southampton (1997)
4. Bray, C., Woodsford, P.: Positional Accuracy Improvement: Impacts of improving the positional accuracy of GI databases. In: EuroSDR official workshop publication. no. 48, Dublin. http://www.eurosdr.net/km_pub/no48/html/positional/positional_index.htm (2004)
5. Roensdorf, C.: PAI2: Achieving Geometric Interoperability of Spatial Data. In: EuroSDR official workshop publication. no. 49, Dublin: http://www.eurosdr.net/km_pub/no49/html/PAI2/index.htm (2005)
6. Hettwer, J., Benning, W.: Restklaffenverteilung mit der Natural-Neighbour-Interpolation. Allg. Vermess.-Nachr. **110**, 122–129 (2003) (in German)
7. Chrisman, N.R.: The Error Component in Spatial Data. In: Maguire, D.J., Goodchild, M.F., Rhind, D. W. (eds.) Geographical Information Systems, Longman, London (1991)
8. Goodchild, M. Gopal, S. (eds.): Accuracy of Spatial Databases. Taylor & Francis, London (1989)
9. Roensdorf, C.: Achieving and Maintaining Interoperability of Spatial Data, FIG working week proceedings, Cairo. http://www.fig.net/pub/cairo/papers/ts_14/ts14_02_ronsdorf.pdf (2004)
10. Straub, B.M., Wiedemann, C.: Towards the update of geodata by automatic object extraction. In: Serpico, S.B (ed.) Image and Signal Processing for Remote Sensing VI. SPIE Proceedings vol. 4170, p. 304. The International Society for Optical Engineering Bellingham, WA (2000) http://ipi216.ipi.uni-hannover.de/html/publikationen/2000/straub/straub_gerke_wiedemann.pdf
11. Saalfeld, A.: Conflation: Automated Map Compilation. US Bureau of Census, SRD Research Report Number: Census/SRD/RR-87/24. http://www.census.gov/srd/papers/pdf/rr87-24.pdf (1997)
12. Kampshoff, S.: First International Workshop on Next Generation 3D City Models. Bonn (2005) http://www.gia.rwth-aachen.de/Service/document_download.php?fileTag=kambonn05
13. Ordnance Survey, Positional Accuracy Improvement website, www.ordnancesurvey.co.uk/PAI
14. MAF/Tiger Accuracy Improvement Program, http://www.census.gov/geo/mod/maftiger.html. This article has been prepared for information purposes only. It is not designed to constitute definitive advice on the topics covered and any reliance placed on the contents of this article is at the sole risk of the reader

Position-Aware Technologies

► Location-Aware Technologies

Positioning

► Location-Based Services: Practices and Products

PostGIS

CHRISTIAN STROBL
German Remote Sensing Data Center (DFD),
German Aerospace Center (DLR), Weßling, Germany

P

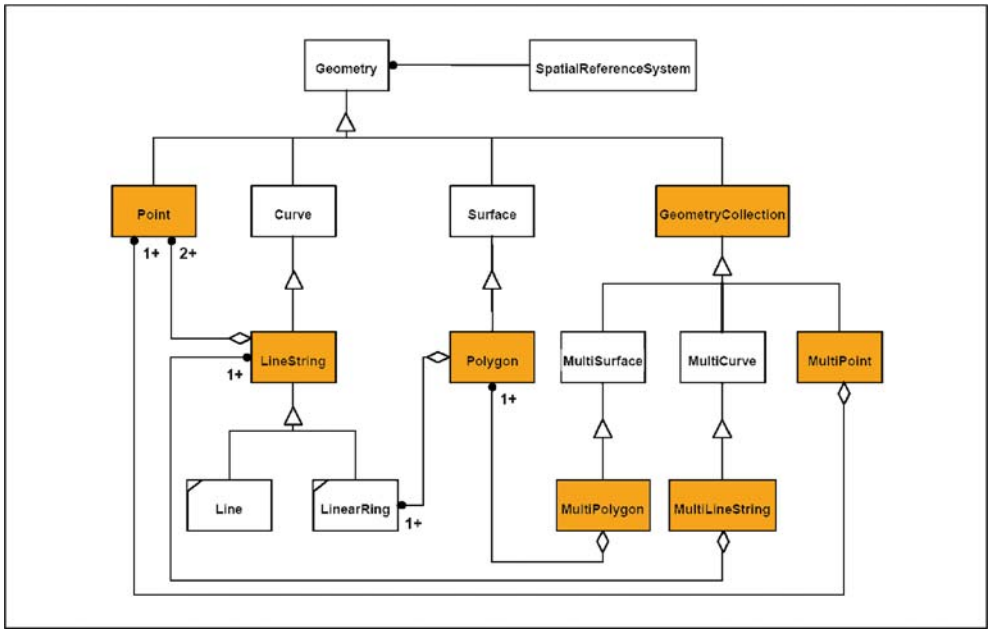
Synonyms

Postgres; OGIS; Spatial DBMS; Public-domain software; Open source; Object-relational; Simple features model; GEOS library; SQL, spatial; R-tree; GiST index

Definition

PostGIS is a spatial database extension for the PostgreSQL (SQL being structured query language) object-relational database. It is certified as a compliant "Simple Features for SQL" database by the Open Geospatial Consortium (OGC).

PostGIS adds geometry data types and spatial functions to the PostgreSQL database. The supported geometry data types are "Points," "LineStrings," "Polygons," "MultiPoints," "MultiLineStrings," "MultiPolygons" and "GeometryCollections". Spatial functions enable the analysis and processing of geographic information systems (GIS) objects. Examples are measurement functions like "Area," "Distance," "Length" and "Perimeter" and spatial operators like "Union," "Difference," "Symmetric Difference" and "Buffer". Topological relationships, like "Equals," "Disjoint," "Intersects," "Touches," "Crosses," "Within," "Contains" and "Overlaps", are processed by the



PostGIS, Figure 1 Geometry class hierarchy of the "Simple Features for SQL" specification from the Open Geospatial Consortium. The geometry types supported by PostGIS are gray shaded

Dimensionally Extended Nine-Intersection Model (DE-9IM).

PostGIS and PostgreSQL are open source. PostGIS is released under the GNU General Public License and PostgreSQL is released under the Berkely Software Distribution (BSD) license.

The functionality of PostGIS is comparable to ESRI ArcSDE, Oracle Spatial, and DB II spatial extender.

Historical Background

The first version of PostGIS was released in 2001 by Refractions Research. It is published under the GNU General Public License [1] and development has continued since then. In 2006, PostGIS was certified as a compliant Simple Features for SQL database by the OGC. It uses libraries of other open source projects. The GEOS (Geometry Engine Open Source) library [2] provides most of the operations described by the OGC Simple Features and the proj4 [3] library contributes the projection support.

Refractions Research is located in Victoria, British Columbia, Canada. It is a consulting and product development organization, specializing in spatial and database application development [4].

The history of PostgreSQL begins at the University of California at Berkeley (UCB). PostgreSQL, originally called Postgres, was created at UCB by a computer science professor named Michael Stonebraker. Stonebraker started Postgres in 1986 as a follow-up project to its predecessor Ingres. Stonebraker and his graduate students actively developed Postgres for 8 years. In 1995, two Ph.D. students from Stonebraker's lab, Andrew Yu and Jolly Chen, replaced Postgres' POSTQUEL query language with an extended subset of SQL. They renamed the system to Postgres95. In 1996, Postgres95 departed from academia and started a new life in the open source world under the BSD license [5]. At the same time the database system was given its current name PostgreSQL. PostgreSQL began at version 6.0 (1996); in 2007 the current version is PostgreSQL 8.2 [6].

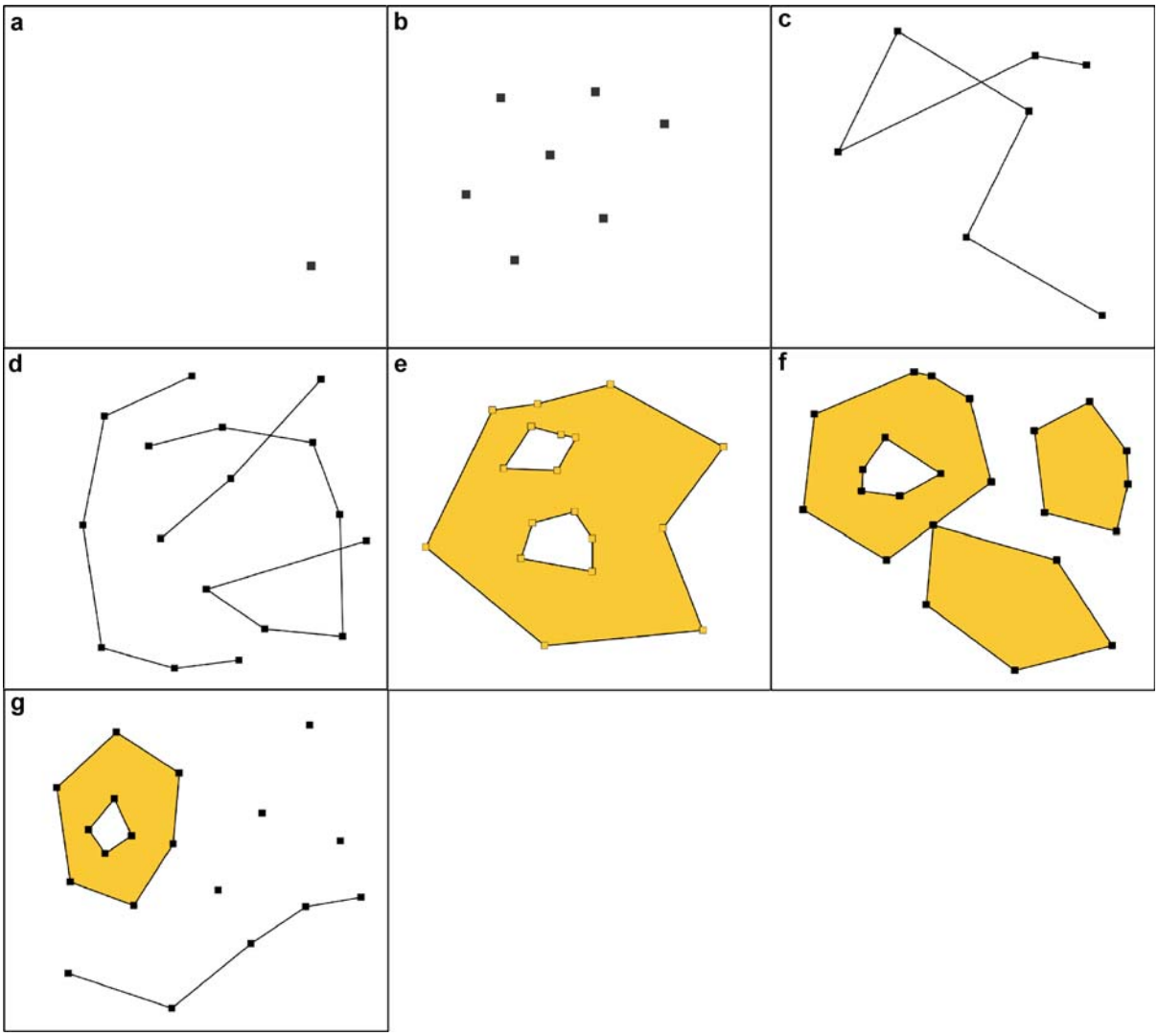
Scientific Fundamentals

Like other spatial databases, PostGIS combines the advantages of classical GIS software, mainly the possibility of spatial analysis, with the advantages of database management systems (DBMS) Such as indexing, transactions and concurrency [7,8].

Simple Features for SQL

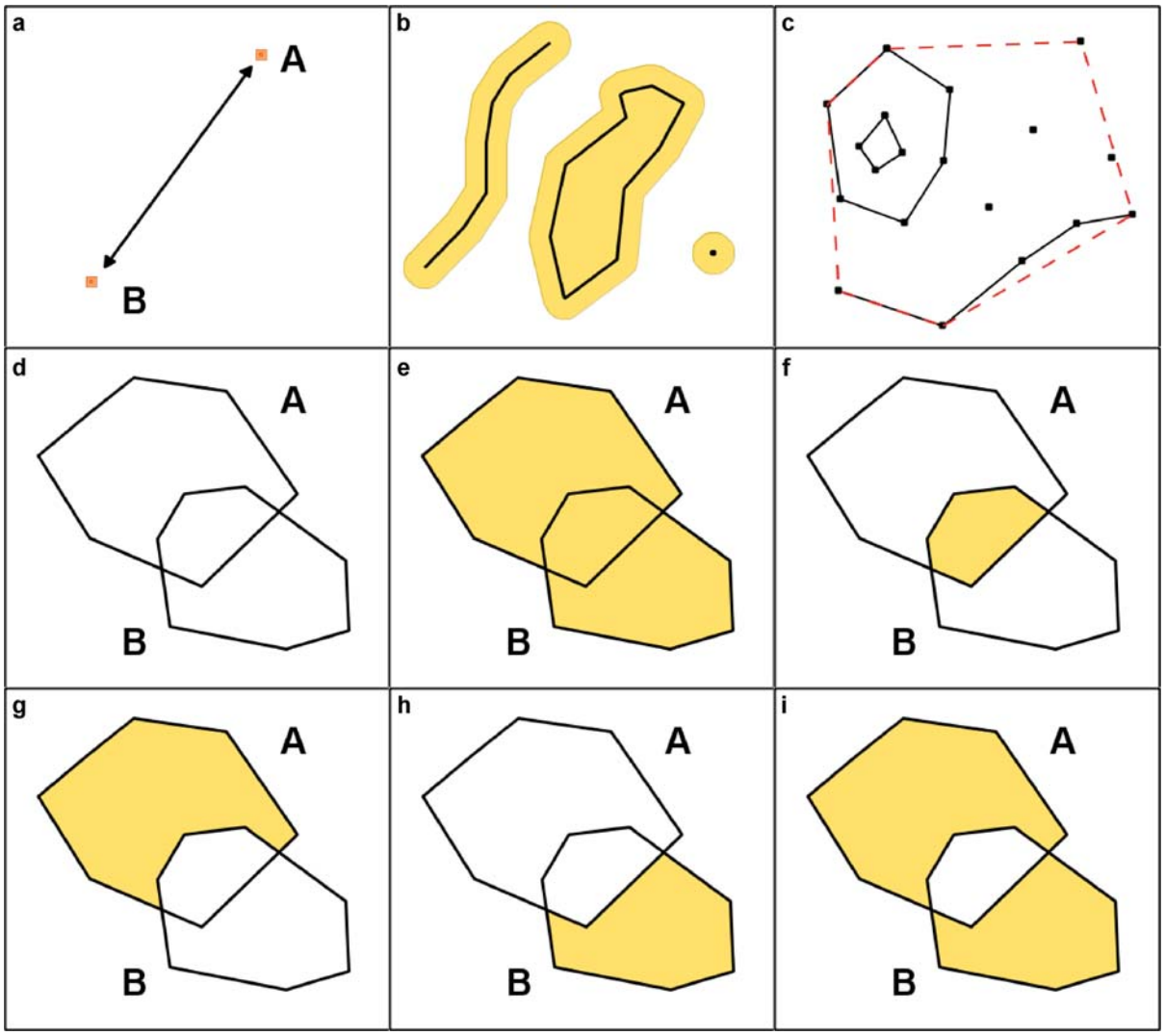
PostGIS follows the Simple Features for SQL specification from the OGC [9]. This implies:

- PostGIS supports the Simple Feature Class Hierarchy according the Open GIS Geometry Model. This includes geometry types for Points, LineStrings, Polygons, MultiPoints, MultiLineStrings, MultiPolygons and GeometryCollections (Fig. 1 and Fig. 2).



PostGIS, Figure 2 Geometry types supported by PostGIS. **a** Point. **b** MultiPoint. **c** LineString. **d** MultiLineString. **e** Polygon. **f** MultiPolygon. **g** GeometryCollection

- PostGIS supports the representation of geometry data as Well Known Text (WKT), Well Known Binary (WKB), as Geography Markup Language (GML) and as Keyhole Markup Language (KML) for Google Earth. Additionally, it supports output as Scalable Vector Graphics (SVG) path geometry.
- PostGIS implements SQL functions that test spatial relationships. These functions include "Equals," "Disjoint," "Intersects," "Touches," "Crosses," "Within," "Contains," "Overlaps" and "Relate". All these operators are based on the DE-9IM [9,10].
- PostGIS implements SQL functions that support spatial analysis. These functions include "Distance," "Buffer," "Convex Hull," "Intersection," "Union," "Difference" and "Symmetric Difference" (Fig. 3).
- PostGIS implements spatial operators for determining geospatial measurements like Area, Distance, Length and Perimeter.
- PostGIS provides information about the geometry type and the spatial reference system. This spatial metadata is stored in the Geometry Columns Metadata View and in the Spatial Reference System Information View according to the Simple Features for SQL specification [9]. Each reference system has a unique identifier called SRID according to the European Petroleum Survey Group (EPSG) code [11].



PostGIS, Figure 3 Spatial functions supported by PostGIS. **a** Distance. **b** Buffer. **c** Convex hull. **e** Union. **f** Intersection. **g** The difference of polygon *A* to polygon *B*. **h** The difference of polygon *B* to polygon *A*. **i** Symmetric difference. **d** The polygons used for the spatial operations of **e**–**i**

Spatial SQL

The implementation of the OGC "Simple Features for SQL" offers GIS new and powerful features for managing, retrieving and analyzing geospatial data. The spatial domain introduces a new set of functions to the SQL Language. The following queries are not complete and give only an elementary review of the potential of the spatial SQL syntax provided by PostGIS:

List the names of all cities which are located inside Bavaria.

```
SELECT city_name
    FROM city a, country b
    WHERE WITHIN (a.geom, b.geom)
    AND b.country_name = 'Bavaria';
```

```
    city_name
---------------------------
Munich
Augsburg
...
```

List the names of all countries which are neighbors to Bavaria.

```
SELECT b.country_name
    FROM country a, country b
    WHERE TOUCHES (a.geom, b.geom)
```

```
        AND a.country_name = 'Bavaria';

          country_name
----------------------------
Thuringen
Baden-Wurttemberg
Hessen
Sachsen
(4 rows)
```

List the names of all cities which are located within 50 km of the river Isar.

```
SELECT DISTINCT a.city_name
    FROM city a, river b
        WHERE DISTANCE (a.geom, b.geom) < 50000
        AND b.river_name = 'Isar';

          city_name
----------------------------
Munich
Passau
...
```

Calculate the area of a buffer of 50,000 m around Munich (see also Fig. 3b).

```
SELECT AREA (BUFFER(geom,50.000))/
                                10000 AS Hectares
    FROM city
        WHERE city_name = 'Munich';

          hectares
------------
780361.288064939
(1 row)
```

List the name, the population and the area of all countries which have an area greater than 3,000,000 ha sorted by the population (in ascending order).

```
SELECT country_name, pop_admin, AREA(geom)/
                                10000 AS Hectares
    FROM country
        WHERE AREA(geom) > 30000000000
        ORDER BY pop_admin;

country_name        | pop_admin | hectares
--------------------+-----------+-----------------
Niedersachsen       | 8000909   | 4733454.20332757
Baden-Wurttemberg   | 10717419  | 3621827.63163362
Bayern              | 12469000  | 7029553.1603014
Nordrhein-Westfalen | 18058000  | 3438200.2301504
(4 rows)
```

Show the geometry type of the table cities.

```
SELECT DISTINCT GEOMETRYTYPE(geom)
    FROM city;

geometrytype
------------
POINT
(1 row)
```

Show the description of the spatial reference system for the table countries.

```
SELECT DISTINCT SRID(geom)
    FROM country;

srid
----
4326
(1 row)
```

With the result of the last query, e. g., the SRID = 4,326, it is possible to get information about the used projection from the table spatial_ref_sys.

```
SELECT srid, proj4text
    FROM spatial_ref_sys
        WHERE srid = 4326;

srid | proj4text
-----+----------------------------------
4326 | +proj=longlat +ellps=WGS84
                +datum=WGS84 +no_defs
(1 row)
```

Show the point location of Munich as WKT.

```
SELECT city_name, ASTEXT(geom) AS "Location"
    FROM city
        WHERE city_name = 'Munich';

city_name | Location
----------+-------------------------
Munich    | POINT(11.5429545454545
                  48.1409727272727)
(1 row)
```

Show the point location of Munich as GML, transformed to the coordinate system with EPSG code 31464 (Gauß Krüger, Germany, 12th meridian).

```
SELECT ASGML(TRANSFORM(geom,31464),7)
    FROM city
        WHERE city_name = 'Munich';

asgml
--------------------------------------
<gml:Point srsName="EPSG:31464">
<gml:coordinates>4466089,5333763
                    </gml:coordinates>
```

```
<gml:Point>
(1 row)
```

Spatial Join (Query Processing)

PostGIS supports spatial joins. A spatial join is comparable to a standard table join based on a spatial relationship. A standard table join merges two tables into one output result. The join is based on a common key.

```
SELECT a.city_name, b.country_name
    FROM city a, country b
        WHERE a.country_name = b.country_name
            ORDER BY b.country_name, a.city_name;
```

A spatial join merges two tables into one output result based on a spatial relationship. For example, the names of the countries are stored in the table country and the names of the cities are stored in the table city. If anybody wants to list the name of the cities and the name of the countries, in which the cities are located, in one table, they have to use a spatial join:

```
SELECT a.city_name, b.country_name
    FROM city a, country b
        WHERE WITHIN(a.geom, b.geom)
            ORDER BY b.country_name, a.city_name;
```

Indexing and Query Optimization

PostgreSQL supports compound, unique, partial, and functional indexes, which can use any of its B-tree, R-tree, hash, or Generalized Search Tree (GiST) storage methods. GiST indexing is an advanced system, which provides an interface and framework for developers to add their own indexes. It allows the combination of a lot of different sorting and searching algorithms including B-tree, B+-tree, R-tree, partial sum trees, ranked B+-trees and others [6,12,13,14].

PostGIS indexes are R-tree indexes, implemented on top of the general GiST indexing schema. R-trees organize spatial data into nesting rectangles for fast searching ([4,15], Fig. 4).

With PostgreSQL and PostGIS, several possibilities exist for query optimization. It is possible to choose between a sequential scan and an index scan for attribute data and between a sequential scan and an index scan using the GiST index for geometry data.

For mixed spatial/nonspatial queries it is possible to use the index with the best selectivity to provide high-performance query plans.

The spatial indexes are not used automatically for every spatial request or operator. Because the R-tree index is based on rectangles, spatial indexes are only efficient for bounding box comparisons. In PostGIS, the indexed search is activated by using the "&&" operator, which means "bounding boxes overlap". The following SQL statement shows a short example:

```
SELECT river_name FROM river
WHERE geom && SETSRID('BOX3D
                    (11 47, 12 49)'::BOX3D,4326)
AND DISTANCE( geom, GEOMFROMTEXT
                    ( 'POINT(12.0 48.5)', 4326) ) < 1;
```

The example query demonstrates a characteristic "two-step" approach to spatial processing:

- The first step, the so-called filter step, is the indexed bounding box search, which runs on the whole table (geom && BOX3D).
- The second step is the so-called refinement step. It only operates on the filtered subset using the exact geometries and represents the original query, in this case the distance query. As this query runs only on the subset returned by the filter step the high costs of processing the exact feature geometries are minimized.

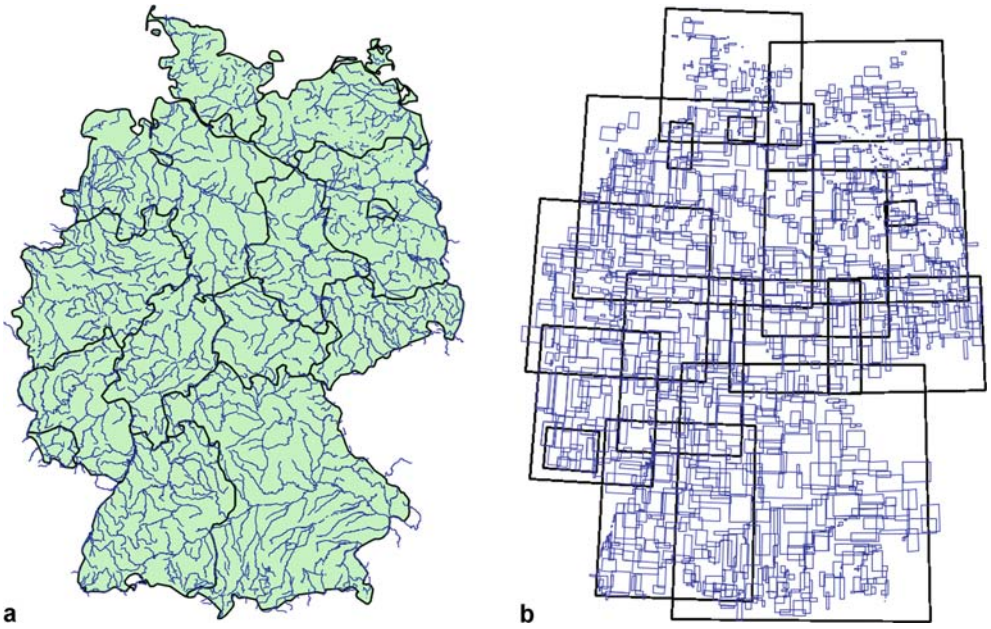
This highly recommended strategy to improve the performance of spatial queries is called the filter-refine paradigm [7].

Key Applications

Spatial data infrastructures (SDIs) facilitate access to geospatial information using a minimum set of standard practices, protocols, and specifications [16]. Every SDI requires a spatial database server and PostGIS represents an open-source- and OGC-compliant solution. Thus PostGIS is supported by many GIS applications, which cover a broad range from server, over workstation and desktop, to internet solutions.

Open Source Software

- deegree: http://www.deegree.org/
- GeoServer: http://geoserver.org/
- GeoTools: http://geotools.codehaus.org/
- GRASS: http://grass.itc.it/
- gvSIG: http://www.gvsig.gva.es/
- MapServer: http://mapserver.gis.umn.edu/
- OGR Simple Feature Library: http://gdal.maptools.org/ogr/
- OpenJUMP: http://openjump.org/wiki/show/HomePage
- Quantum GIS: http://www.qgis.org/
- Thuban: http://thuban.intevation.org/
- uDig: http://udig.refractions.net/
- ...



PostGIS, Figure 4 The bounding boxes that are used for the spatial indexes of the countries and rivers shown in **a** are given in **b**

Proprietary/Closed Software

- ArcGIS (with the Interoperability Extension): http://www.esri.com/
- Cadcorp SIS: http://www.cadcorp.com/
- Feature Manipulation Engine FME: http://www.safe.com/
- Ionic Red Spider: http://www.ionicsoft.com/
- ...

Future Directions

The 1.2.0 release of PostGIS comes with the first support for "curve" types, based on the International Organization for Standardization (ISO) SQL/MM (SQL Multimedia and Application Packages) model for curves. Also initial support for the ISO SQL/MM suite of spatial database functions is implemented [17].

In addition to the ongoing implementation of the ISO SQL/MM standard the PostGIS team works on three-dimensional surface and spline curve support, topology, networks, routing, long transactions and raster integration. The initial groundwork for using PostGIS as an ESRI ArcSDE style interface was also laid in version 1.2. This includes support for most of the ST_* and SE_* spatial SQL functions used by the ArcSDE spatial SQL interfaces.

Cross References

- ▶ Data Infrastructure, Spatial
- ▶ deegree Free Software
- ▶ Dimensionally Extended Nine-Intersection Model (DE-9IM)
- ▶ OGC's Open Standards for Geospatial Interoperability
- ▶ Oracle Spatial, Geometries
- ▶ University of Minnesota (UMN) Map Server

P

Recommended Reading

1. Free Software Foundation (1991) GNU General Public License. http://www.gnu.org/copyleft/gpl.html. Accessed 1 Jan 2007
2. GEOS (2006) Geometry Engine Open Source. http://geos.refractions.net. Accessed 1 Jan 2007
3. PROJ.4 (2000) Cartographic Projections Library. http://www.remotesensing.org/proj. Accessed 1 Jan 2007
4. Refractions Research (2007) PostGIS. http://postgis.refractions.net/. Accessed 1 Jan 2007
5. University of California (Berkeley) (1999) The BSD License. http://www.opensource.org/licenses/bsd-license.php. Accessed 1 Jan 2007
6. PostgreSQL Global Development Group (2007) PostgrSQL. http://www.postgresql.org/. Accessed 1 Jan 2007
7. Shekhar, S., Chawla, S.: Spatial Databases: A Tour. Prentice Hall, Upper Saddle River, NJ (2003)
8. Rigaux, P., Scholl, M., Voisard, A.: Spatial Databases: With Application to GIS. Morgan Kaufmann, San Fransisco (2002)
9. OGC (ed.): OpenGIS Simple Features Specification for SQL (Revision 1.1). OGC (1999)
10. ISO/TC211: ISO 19107 Geographic Information—Spatial Schema. ISO (2003)
11. OGP Surveying & Positioning Committee (2006) EPSG Geodetic Parameter Dataset v 6.11.2. http://www.epsg.org. Accessed 1 Jan 2007

12. The GiST Indexing Project (1999) GiST: A Generalized Search Tree for Secondary Storage. http://gist.cs.berkeley.edu. Accessed 1 Jan 2007
13. Sigaev, T., Bartunov, O. (2006) GiST for PostgreSQL. http://www.sai.msu.su/~megera/postgres/gist. Accessed 1 Jan 2007
14. Hellerstein, J.M., Naughton, J.F., Pfeffer, A.: Generalized search trees for database systems. In: Proceedings of the 21st International Conference on Very Large Data Bases, Zürich, Switzerland, 11–15 Sept 1995
15. Guttman, A.: R-trees: a dynamic index structure for spatial searching. In: Proceedings of ACM SIGMOD International Conference on Management of Data Boston, MA, 18–21 June 1984
16. Nebert, D.D. (2004) Developing Spatial Data Infrastructures: The SDI Cookbook, Version 2.0. http://www.gsdi.org/pubs/cookbook/cookbookV2.0.pdf. Accessed 12 Oct 2005
17. ISO/IEC: ISO/IEC 13249-3: Information Technology—Database Languages—SQL Multimedia and Application Packages. Part 3: Spatial (2006)

Postgres

▶ PostGIS

Preference Structure

▶ Multicriteria Decision Making, Spatial

Prism, Network Time

▶ Time Geography

Prism, Space-Time

▶ Time Geography

Privacy

▶ Cloaking Algorithms for Location Privacy

Privacy and Security Challenges in GIS

BHAVANI THURAISINGHAM, LATIFUR KHAN, GANESH SUBBIAH, ASHRAFUL ALAM, MURAT KANTARCIOGLU
Department of Computer Science,
The University of Texas at Dallas, Dallas, TX, USA

Synonyms

Geographic data management

Definition

Geospatial data refers to information about shapes and extent of geographic entities along with their locations on the surface of the earth. This definition, however, is often extended to include any physical or logical entity as long as it exhibits one or more geographic characteristics such as topology of a proposed highway infrastructure or location of a moving vehicle. Geospatial data management pertains to the acquisition, manipulation and dissemination of geospatial data under a set of guidelines. It has numerous applications including counter-terrorism, climate-change detection and space exploration. For example, global warming has been one of the major climate changing events in recent years. The significance of global warming lies in the severe impact that even small climate changes could cause on weather patterns, ecosystems and other activities. Understanding the causes and impacts of global warming is therefore critical. Central to this mission are the thousands of stations capturing vast amounts of geospatially referenced climate and weather data, both on and off the Earth. The data is stored in hundreds of geographically distributed databases, often in different formats. Even more problematic is that the data lack a common semantics, and as a result tends to take on different meanings in different places. These two problems are major impediments to scientists in their ability to coherently and consistently analyze the data, and investigate global trends, make predictions, and so forth.

One way to effectively analyze and detect climate changes is to apply knowledge discovery techniques, also referred to as data mining, for geospatial data sources. If the experts are to systematically process the data in order to answer important scientific and social questions, a coherent representation of the geospatial data related to global warming is needed. The semantic heterogeneity problem is handled by establishing domain ontologies (e. g., emission model, temperature model, sea-level model) to aid in the process of data annotation. A large number of existing environmental parameters can be mapped to geospatial data objects and the remaining ones could be added on gradually.

While the geospatial data related to climate modeling and changes, as well as much of the geospatial data such for counter-terrorism applications such as photographs of building and bridges, are usually publicly available, certain fields may be sensitive to a particular organization. Furthermore, the results of the integration and analysis of the geospatial data may also be sensitive. A recent report by Rand Corporation has stated that geospatial data, even those publicly available, have security needs that must be dealt with [1]. National Oceanic and Atmosher-

ic Administration (NOAA) has also discussed the strong need for security policy enforcement for climate data records (CDR) [2].

Much progress has been made on geospatial information systems such as the specification of the geospatial markup language (GML) [3] for data representations by organizations such as the Open Geospatial Consortium (OGC) as well as information retrieval techniques. However several areas including techniques for integrating geospatial data as well as mining the data needs research. Furthermore, security and privacy issues have received very little attention for geospatial data management integration and mining. Research in the areas of geospatial data integration, mining and security are being conducted.

Historical Background

Some past research work has been reported on secure geospatial data management systems [4,5], as well as secure web services and secure semantic web [6]. For example, Atluri [4] has proposed a model that takes into account the characteristics of geospatial data. Bertino et al. [7] have developed a model called GEORBAC that extends role-based access control (RBAC) for geospatial data that take into consideration classification policies depending on content, content and time. The OGC members have also done some exploratory work in the use of Public Key Infrastructure (PKI) and extensible access control markup language (XACML) for building and deploying more secure geospatial portal applications. The OGC is also working on standards for geospatial digital rights management. However, in the literature survey done there is no work on developing secure geospatial semantic web and web services except for the research being conducted at the University of Texas at Dallas [8].

In a service-oriented architecture or a distributed system where multiple parties collaborate to exchange geospatial data, it is imperative that a strong security mechanism is maintained to ensure participating parties' continued willingness to share data. The abundance of data exchange protocols and the varying business needs of the parties make it a challenging task to devise an appropriate security model. The security specification from the Organization for the Advancement of Structured Information Standards (OASIS) defines a web service security model that unifies several popular security models and technologies to be able to interoperate in a platform- and language-neutral manner. XACML is the OASIS security standard, which allows developers to write and enforce information access policies for web services. The web service policy language (WSPL) is another proposed language for web services security framework. These languages lack inference and reasoning capabilities as they are not semantics-aware frameworks for machines to interpret, although they establish syntactical interoperability. GeoXACML [9] is an access control language proposed for geospatial web services.

There are two overlooked aspects in the existing security models mentioned above. First, they are mainly suitable for a single-party environment. In an integrated environment where resources come from various parties, the individual policies of each party have to be combined to apply in a global context. Bertino et al. [10] have proposed an integration algorithm for combining access policies of multiple autonomous parties in a distributed environment. They extend XACML by including a set of preferences that allow dynamic computation of policy integration need. The other overlooked aspect in the current models is the lack of semantics awareness in policy constructs. Semantic Web allows a platform for policy reasoning and inferencing if the policies are written in a semantic-aware language. Although the techniques for Semantic Web security are yet to be standardized, there has been work involving security ontologies. Different policy representations have been proposed using semantic languages such as Rei, and KAoS. KAoS exploits ontologies for representing and reasoning about domains describing organizations of humans and agents. Rei is a deontic concept-based policy language in Resource Description Framework-Schema (RDF-S).

One of the major challenges confronted in geospatial data management is collection and assimilation of data without major loss of fidelity. The most commonly employed approach has been using geospatial systems or ad-hoc programs to define methods that convert data from one source or format to another with the help of wrappers. This approach has limitations in so far as the wrappers are cumbersome and require manual translations every time a new data format or standard appears. Several proposals have been offered that utilize schema mechanisms (e. g., GML) to define concepts in a standardized manner. Nonetheless, the semantics provided by the schemas for geospatial resources are not machine-readable and hence are difficult to share between systems without prior coordination. While there have been researches to address these limitations (e. g., [11]), a comprehensive approach to developing a geospatial semantic web with appropriate technologies for specifying semantics as reasoning engines are yet to be developed.

Scientific Fundamentals

There are different levels of interoperability issues that need to be addressed when two or more geospatial data sources are to be integrated. One of the major problems

is semantic heterogeneity. An example is the following: land cover classifications where definitions of forest, plantation, wood, copse, scrub, orchard, etc. all relate to areas with some tree cover but different organizations and countries may use them differently as well as use different terms for the same entity. The other problem is structural heterogeneity. For instance, a geographic location can be expressed by, for example, a closed string, and two separate coordinates or by a point. Research on semantic interoperability between geospatial data sources of the same theme is underway. A major challenge is to integrate the work of OGC and the World Wide Web Consortium (W3C) to develop a geospatial semantic web that handles semantic heterogeneity. Another challenge is in the development of geospatial semantic web services that can discover and manage resources in a global environment.

While integration of data sources is important, it has to be done securely to ensure participating parties' willingness in sharing their data. An important security consideration in this process is the integration of security policies. Since the individual agencies implement their own security policies to protect the data, several critical issues arise during the policy integration. The first issue is the mismatch of policy rule semantics. That is, when a policy has to be integrated with other policies, attributes and targets of the policies should be interpreted consistently by the system. For example, if two policies from separate agencies use "manager" and "supervisor" respectively, to specify the same role attribute, the integration algorithm should be able to interpret this equivalency. The second issue is rules mismatch. Even if the assumption of no heterogeneity is made, attributes sets and targets of separate policies have to be matched properly.

Further security challenges include coming up with appropriate policies for climate and weather data, as well as language to specify the policies. Policies may depend on content, context and time. Different agencies may enforce different policies. Furthermore, collections of data from multiple databases within an agency or from multiple agencies taken together may be sensitive, while individually they may not be classified. The geospatial semantic web is expected to provide a level of semantics to help in designing secure contextualized and georeferenced policies that reason about their robustness.

A study was conducted to evaluate existing geospatial web service standards against the requirements identified in the use cases, in particular, identification of formal change requests to enhance existing standards. In those cases where existing standards will not work or cannot be adapted, identifying and developing new web service interface standards was investigated. In both cases, the focus was on (1) geospatial semantic web services for applications such as discovering and managing geospatial data resources, and (2) geospatial semantic web technologies for information integration and related security considerations. Both types of services are closely intertwined as the information integration application will invoke the geospatial semantic web services for providing various services. Each web service has a high-level service description that is written using Web Ontology Language for Services (OWL-S). OGC specifies geospatial interface and encoding standards. The key encoding standard is the GML. OWL-S provides a semantic rich application level platform to encode the web service metadata using descriptive logic. The approach used is essentially the following:

- Semantic enrichment of the OGC web services framework by using OWL-S ontology
- Query disambiguation of the service requestor using semantics
- Automatic service discovery and selection using capability-based matchmaking
- Automatic service composition and invocation

Since geospatial data involves geospatial constructs such as overlap and boundary which are required to be disambiguated during the query phase, the registered services will then be automatically discovered for the disambiguated query using capability-based search which is a more expressive mechanism than the simple keyword-based search currently used in the service registries. The selected web services will be automatically invoked using the WSDL groundings. Dynamic service compositions on the fly are made possible for the service requestor's query. The research carried is for developing geospatial semantic web technologies for information integration. Development of a geospatial resource description framework (GRDF) that extends GML to include semantics [8] has been initiated. This is also intendend to enhance GRDF (e. g., extensions to support climate data), which is the foundation for a geospatial semantic web, and subsequently extend the reasoning engines (such as those in JENA; JENA is a java framework for building semantic web applications) for geospatial data.

Research is also needed to investigate security issues for geospatial semantic web and web services. The core of the approach is represented by a semantically rich web service access control model consisting of a policy layer that processes user queries to geospatial web service agents. The security policies have to be enforced and only the authorized data is retrieved and returned to the user. In the case of multiple geospatial data servers, each node may enforce its own set of policies as specified and enforced by the policy framework. Data access by a web service is mediated by a broker and the request is then sent to different locations. Since policy descriptions and granularity will

be annotated in descriptive logic (i. e., OWL-DL), the proposed access control model will allow automatic reasoning between communicating clients and agents. A secure GRDF language is being developed examined to specify the security semantics.

There are unique challenges for discovering knowledge from climate-change-specific geospatial data. For example, Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) data is used to model detailed maps of land surface temperature, emissivity, reflectivity, and elevation. This characteristic of ASTER data offers opportunities to observe, understand, and model the Earth system, enabling us to better predict change, and to understand the consequences for life on Earth. ASTER obtains high-resolution (15–90 m^2 pixel) images of the Earth in 14 different wavelengths of the electromagnetic spectrum, ranging from visible to thermal infrared light. Therefore, there is a need to build higher-level knowledge from this data for analyzing complex phenomena. Geospatial data specific to climate change has spatial and temporal characteristics that add substantial complexity to data-mining tasks. The spatial relations, both metric (such as distance) and nonmetric (such as topology, direction, shape, etc.) and the temporal relations (such as before and after) are information bearing and therefore need to be considered in the data mining methods.

Data mining raises serious security and privacy concerns. There are two aspects here; one is that the results of data mining may be sensitive. The other is that while the individual climate data records are sensitive, the results of the data mining tool are unclassified. These issues have to be investigated for geospatial data.

For climate change, current work focuses mainly on the change detection of various classes (i. e., "urban area", "forest" and so on) that appear in images of a particular location over time. The tasks involved for such an approach include identifying the class/label of pixels in images, estimating contiguous areas in the map/image that belong to the same class, and comparing areas of the same class taken from two different images for the same location and determining changes. For example, from 1986 to 1998, urban areas increased a total of 52,019 ha or by 28.4%. This number can be estimated by first classifying urban areas in images for 1986 and 1998 separately and then estimating the difference. To classify pixel value into various classes, the current state-of-the art uses a maximum likelihood (ML) classifier; it has been observed that the accuracy of ML is not satisfactory. Lower accuracy may contribute higher false positives and higher false negatives for climate change detection [12].

As far as the authors know, security for geospatial data mining has not received any attention. At the University of Texas at Dallas research has started in this field with respect to both confidentiality and privacy.

The approach consists of the following:

- Extracting features to facilitate climate change detection
- Training classifiers using extracted features and predicting class/label of pixels that appear in images
- Comparing contiguous areas of the same class taken from two different images for the same location to facilitate change detection
- Correlating these atomic concepts/classes to make a decision of generic concept with the help of ontologies

For feature extraction, ASTER data has 14 channels, from visible through the thermal infrared regions of the electromagnetic spectrum, providing detailed information on surface temperature, emissivity, reflectance, and elevation. ASTER provides valuable scientific and practical data of the Earth in various fields of research. To classify pixels that appear in images, research is by exploiting various data mining techniques including support vector machines (SVM) combined with a developed technique called Dynamically Growing Self Organizing Tree (DGSOT) [13]. Investigation has shown that SVM+DGSOT is a powerful method for classification. This classifier will help to determine atomic classes/concepts. Change detection can be done by comparing contiguous areas of the same class taken from two different images for the same location. Exploiting ontologies with embedded rules will enable the determination of generic concept/outcome. For example, a set of high-level concepts (i. e., wildfire) can be inferred using ontologies and a set of atomic concepts (e. g., low rainfall). In particular, exploiting ontology-based concept learning improves the accuracy of the individual concept. This is achieved by considering the possible influence relations between concepts based on the given ontology hierarchy.

Two aspects with respect to security need examination. First, the prior research on enforcing security and privacy constraints for data management systems must be examined, and the inferencing techniques for classifying the results produced by the data mining tools applied. Previous work in secure multiparty-based cryptographic approaches for privacy preserving data mining as well as other approaches should also be examined, and techniques developed for security/privacy preserving geospatial data mining [14].

Key Applications

Geospatial data are becoming increasingly useful across many different applications for enhancing the visual aspect

of the raw data and providing additional dimensions to enable decision making and analysis. Some of the most promising and critical applications are described here.

Emergency Response System

In the case of an emergency, first responders and decision-making personnel often need to gather and analyze georeferenced data on the fly. Without efficient data management, collecting and presenting the pertinent data in a coherent form would be unfeasible.

Climatology

Geospatial data includes information regarding weather patterns, seasonal changes, wind velocity, and atmospheric and sea-level pressure and so on. Proper collection and filtering of this data is critical in studying climate trends. Climate changes that are deviating from the norm or that imply serious repercussions can be determined based on the collected data.

Semantic Web

Semantic web refers to a distributed system where all kinds of data stores and client applications are connected via a framework that incorporates a loose data model, logic, rules and reasoning. The basic idea behind semantic web is to enable a minimum human-intervention infrastructure and maximum machine automation. The applications on the semantic web can tap into various data sources to fetch the pertinent data, and then merge them to present coherent and precise results to application users. For instance, a semantic-web-enabled automated restaurant finder agent can extract restaurant data and georeferenced data to present not only the route to the destination, but the weather and crime rate in the area as well.

Future Directions

This paper has provided an overview of geospatial data management and discussed the need for security;, geospatial data integration, geospatial data mining and the impact of security and privacy on these functions have been discussed. For each of the functions, challenges have been identified, along with the state of the art and research directions.

As stated earlier, security and privacy are important considerations for geospatial data mining. Even through much of the geospatial data is publicly available, according to the Rand report there are many attributes that have to be protected. Furthermore, the privacy of the individuals has to be maintained. There is still much work to be done in geospatial data interaction, mining, security and privacy.

Acknowledgements

Special thanks go to Dr. Mike Jackson of the University of Nottingham and the Open Geospatial Consortium for their comments on our research. Prof. Elisa Bertino and Prof. Michael Gertz also provided valuable comments on geospatial data security.

Recommended Reading

1. Assessing the Homeland Security Implications of Publicly Available Geospatial Information. Rand Report for NGA (2004)
2. National Research Council: Climate Data Records from Environmental Satellites: Interim Report, NOAA Operational Satellites. National Research Council (2004)
3. Geography Markup Language (GML) Version 3.1.1. http://portal.opengeospatial.org/files/?artifact_id=4700
4. Atluri, V., Chun, S.: An authorization model for geospatial data. IEEE Trans. Dependable Sec. Comput. **1**, 238–254 (2004)
5. Geospatial Interoperability Reference Model (GIRM, V 1.1) http://gai.fgdc.gov/. Accessed December 2003
6. Lieberman, J., Pehle, T., Dean, M.: Semantic Evolution of Geospatial Web Services. http://www.w3.org/2005/04/FSWS/Submissions/48/GSWS_Position_Paper.html
7. Belussi, A., Bertino, E., Catania, B., Damiani, M.L., Nucita, A: An authorization model for geographical maps. http://www.informatik.uni-trier.de/~ley/db/conf/gis/gis2004.html GIS 2004
8. Ashraful, A., Thuraisingham, B.: Geospatial resource description framework (GRDF) and secure GRDF. Tech. Rep. UTDCS-03–06, University of Texas at Dallas (2006) http://www.cs.utdallas.edu
9. Matheus, A.: Declaration and enforcement of fine-grained access restrictions for a service-based geospatial data infrastructure. In: Proceedings of the10th ACM Symposium on Access Control Models and Technologies, pp. 21–28. ACM, New York (2005)
10. Mazzoleni, P., Bertino, E., Crispo, B., Sivasubramanian, S.:XACML policy integration algorithms:~not to be confused with XACML policy combination algorithms! In: Proceedings of 11th ACM Symposium on Access Control Models and Technologies, pp. 219–227. ACM, New York (2006)
11. Li, D.: Geospatial Semantic Web Research at LAITS. http://www.ncgia.ucsb.edu/projects/nga/docs/Di_Position.pdf
12. Buttenfield, B., Gahegan, M., Miller, H. Geospatial Data Mining and Knowledge Discovery. http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emerging/gkd.pdf
13. Khan, L., Awad, M., Thuraisingham, B.: A new intrusion detection system using support vector machines and hierarchical clustering. VLDB J. (2007, in press)
14. Kantarcioglu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. Knowl. Data Eng. **16**, 1026–1037 (2004)
15. Onsrud, H.J., Johnson J.P., Lopez, X.: Protecting personal privacy in using geographic information systems. Photogramm. Eng. Image Process. **60**:1083–1095 (1994)

Privacy of Location Traces

► Privacy Preservation of GPS Traces

Privacy Preservation of GPS Traces

MARCO GRUTESER, BAIK HOH
Electrical + Computer Engineering (WINLAB),
Rutgers University, Piscataway, NJ, USA

Synonyms

Privacy of location traces; Anonymization of GPS traces; Privacy preserving in location based services

Definition

Techniques to evaluate and enhance the privacy of users that contribute traces of their movements to a geographical information system (GIS). Due to the increasing prevalence of global positioning system (GPS) chips in consumer electronics and advances in wireless networking, GIS can collect *GPS traces* of large numbers of individual users. These traces give rise to privacy concerns, since GPS traces can reveal visits to sensitive or private places (e. g., home, medical clinics) and associated information such as time of day or speed of travel.

Privacy can be enhanced through standard data protection techniques such as policy disclosure, obtaining user consent, access control, and encryption. Anonymization, another standard technique, is of particular interest for GIS applications that aggregate data from many users, since it also protects against accidental and insider data breaches and enables public release of aggregate datasets. Anonymization of location traces, however, poses special challenges since the detailed time-series nature of a GPS trace often allows re-identification of users. For example, it is often straightforward to identify the home or work positions based on a GPS trace, providing means to re-identify the user. Thus, removing identity information from a trace only provides weak anonymity. To obtain a strong degree of anonymity, special location disclosure control algorithms must be used that reduce resolution, omit especially sensitive parts of the trace and divide traces into shorter disjoint parts to prevent extended tracking.

Historical Background

Privacy concerns due to the misuse of new technological inventions can be traced back at least to Louis Brandeis article "The right to privacy" [4] addressing photography in 1890. Since then technological advances have posed repeated challenges and have given rise to new social norms, legal concepts, and technological solutions. Early technological solutions for data privacy include data encryption for communication and storage, operating system and database access control, and auditing. Over the past few decades, as information technology has permeated our lives, several notable privacy technology developments occurred that have influenced the development on techniques for GPS traces:

- *Statistical Databases.* These databases allow queries that retrieve aggregate statistics about individuals' data but do not allow retrieval of any individual's record. The main challenge in the development of such databases is protection against inferences that reconstruct individual records from the results of several carefully selected aggregate queries.
- *Privacy-Aware Data Mining.* The objective of these techniques is similar to statistical databases in that they seek to protect individual database records. Instead of calculating precise statistics, these algorithms only allow reconstruction of the approximate distribution of attributes over the total population, which is sufficient for many data mining tasks. For example, privacy techniques, such as value-class membership or time-series distortion, can increase privacy for individual records, while still allowing classifications algorithms to operate on aggregate data [1].
- *k-Anonymity.* The k-anonymity concept provides a formal model for evaluating privacy protection of a dataset. Samarati and Sweeney [14] developed this concept and an algorithm that can remove or generalize sensitive data so that a user's record is indistinguishable from at least $k-1$ other records. Thus, this algorithm enables anonymizing a database table, so that the table can subsequently be released to external sources (e. g., releasing medical records to researchers). Anonymity-based solutions were also developed for enhancing communication privacy [5].

These concepts provide a foundation for the development of location privacy techniques described in the following sections, which were motivated by the advent of affordable positioning and tracking technologies.

Scientific Fundamentals

A typical GPS trace contains a collection of individual position samples, each comprising latitude, longitude, timestamp, and optionally speed and heading. Privacy of a dataset of such traces may be protected through well-known data protection techniques, such as encryption and access control. These techniques are effective, when only a limited number of fully trustworthy users require access to the dataset. The dataset can then be protected from eavesdroppers or curious other users by encrypting the dataset before communication or storage, for example.

Anonymization techniques may be more appropriate if the dataset must be released to a larger number of not fully

trusted parties or when the identities of data providers are not needed for the application. The remainder of this entry will concentrate on such techniques, since the techniques used in this case a more specific to GPS Traces.

A first step towards effective anonymization is removing explicit user identifiers, such as names or cell phone identifiers, that may be associated with the trace. We refer to this as a (weak) anonymous trace.

The exact privacy implications of such anonymous GPS traces depend on many factors, especially GPS accuracy, building density, sampling frequency, trace duration, user density, and other data associated with each sample. First, consider a trivial anonymous GPS trace containing only a single position (latitude, longitude, timestamp). This trace could pose a location privacy risk, if an adversary can infer the user's identity. Identification is possible through

- *Restricted space correlation.* A restricted space is a geographic area only accessible to one known person, such as a home or office. If samples originate from this location, the adversary can infer with high probability the user identity associated with this sample.
- *External observation correlation.* An external observation is a sighting of a single known individual at a give position and time through other means (e. g., electronic toll booth records, credit card transaction data, video surveillance tape, etc.). If no other uses were present at the location and the position and time of the sighting matches the GPS sample, the user can be identified.

Both methods require that some information about the position of the individual is already known. Still, privacy risks can exist when disclosing the GPS traces for three reasons. First, an adversary may learn more precise information about the whereabouts of the individual, for example the exact time an individual was present or the exact room a person visited. Second, the usage of GPS samples poses a more general *data privacy* risk, if other sensitive information is associated with the sample. For example, a user might conduct an apparently anonymous location-based search for the closest medical clinics on a cell phone, which sends a GPS sample associated with the search terms to an external search service provider. If one of the above identification methods is possible, an adversary with access to the service provider logs may connect the search terms (e. g., a medical condition) with a particular individual by using the GPS sample. Third, the traces may reveal information about other visits and activities, if they contain multiple samples. We will discuss this case further below.

Generally, identification through unrestricted space and external observation correlation is feasible, if correlation data is available and the GPS resolution is high enough to uniquely identify a person or space from the correlation data. Modern GPS receivers typically achieve sub 10m accuracy in open-sky areas, enough to uniquely identify most suburban homes, but rarely sufficient to pinpoint an apartment in an urban high-rise building. A similar relationship between GPS accuracy and user density exists for observation identification.

Spatial cloaking [8] provides a countermeasure against these risks. It dynamically adjusts the resolution of position samples to maintain a constant degree of privacy in situations with different user densities. Given a set of traces from different users, the spatial cloaking algorithm achieves k-anonymity by determining a square that encloses the current positions of at least k users. Square corners are chosen from an external reference grid, so that they do not reveal any clues about current user positions. The position samples of the k users are then replaced with the square (or its center point).

The privacy risks for single positions are compounded for longer GPS traces, which contain more than one position sample. If a user can be identified at any one point, an adversary can infer which buildings (e. g., stores, clubs, medical clinics, entertainment venues) a person visited and accurately measure time spend at work or at home. If the frequency of location samples is high (at least one every few minutes) one may also infer speed limit violations while driving, for example, even if the GPS device does not report speed information. Further identification risks are higher, because a person could now be identified through knowledge about the frequency of their visits to each location in the trace [11].

A countermeasure against these particular trace risks is *path segmentation* [2,9,15], which divides several anonymous traces into shorter traces, or in the extreme, into a set of anonymous samples. Intuitively, this might reduce the risks to those identified for anonymous samples. However, an adversary may frequently be able to reconstruct the complete traces by "following the footsteps" (if one segment begins where another one ends the trajectory of both points into the same direction, they likely belong together). This can be automated through location tracking algorithms that exploit the spatio-temporal correlation between subsequent samples, such as *multiple target tracking* [9,13]. In essence, these algorithms predict a user's next position based on the previous trajectory and add the sample closest to the prediction to the trace. This approach fails, if many potential users are near the predicted position—thus, the segmentation approach is only effective in areas where user density is high and many users share common paths. Note that the target tracking algorithms can also filter noise from the location samples, thus privacy techniques that add random noise to each sample may be ineffective, unless the noise compo-

nent is very large compared to the range of possible positions.

Better privacy protection for GPS traces can also be provided through special disclosure control algorithms such as *origin-destination cloaking* (ODC) [10] or *uncertainty-aware path cloaking* [12]. ODC is designed for GIS applications that primarily require applications from moving users, such as traffic monitoring applications in the automotive domain. ODC cloaking aims to suppresses the parts of location traces that are close to locations that a user has visited, but allow release of location information when the user is moving. The intuition behind this approach is that visited locations provide likely avenues for identification and reveal potentially sensitive information. With ODC the exact visited building remains hidden, only the general area is known. Thus, both restricted space identification and compiling a dossier of visited locations becomes more difficult. Uncertainty-aware path cloaking further limits the tracking time when moving by dropping samples when extended tracking was possible.

Key Applications

Many pervasive or context-aware computing applications rely on the availability of periodic and accurate location information provided by ever more cost-effective GPS chips. Applications such as the following that make GPS traces available to external service providers can benefit from the described data privacy techniques:

- Traffic monitoring applications: Instead of camera or loop detectors on the roads, probe vehicles, which are equipped with GPS and sensors, are expected to be used in many traffic monitoring systems [11]. Usage of the described privacy techniques could better protect privacy and increase participation rates in such schemes.
- "Pay as you drive" insurance: This approach allows auto insurance carriers to customize insurance premiums to individual driving patterns. In return for potential discounts, drivers let the insurer install a GPS device that provides GPS traces to the insurer. To improve risk assessment the insurer can then analyze the traces for mileage driven, roads taken, speed, time of day for trips, duration of rest periods, and other factors. While this application likely requires drivers to identify themselves to the insurance provider, techniques like OD cloaking may also be beneficial in this scenario, to reduce the amount of information collected.
- Electronic toll payment: Some next-generation electronic toll collection systems use GPS to calculate more fine-grained distance and time-based tolls. Current radio-tag based systems (e. g., EZ-Pass in New Jersey and FasTrak in California) have been regarded with suspicion since the history of road usages are collected with identity, this lets authority clearly see where subscribers are driving. However, recent research [3,6] proposed more privacy-aware toll collection protocols.
- Cell phone location-based services: Many US cell phone handsets incorporate GPS chips that can provide very precise position information in many cases. These are used primarily to satisfy the E911 regulatory requirements, which mandate that cell phone service providers must be able to locate emergency callers. This infrastructure, however, is also being used for offering location-based services, such as point-of-interest queries or navigation. Spatial cloaking allows users to use these services with enhanced privacy.

Future Directions

Applications that have access to private GPS traces from large numbers of users are relatively new. Thus this area provides many topics for further research.

- *Risk analysis and privacy metrics.* To date little practical experience with such applications exist. Privacy risks are typically identified by studying analogies to risks in other information systems. Improved privacy frameworks and metrics are needed to guide analysis of privacy risks in applications. These frameworks should include quantitative guidance on parameters such as user density, sampling frequency and trace duration.
- *Usable privacy preferences* Since increased privacy protection usually reduces the quality of service provided by the application, a complete privacy solution should allow users to choose or specify different disclosure options. This requires research on user interfaces to understand how users can best express these preferences. It also requires research in privacy algorithms that must remain secure even if some users disclose more detailed information than others.
- *Maintaining privacy when using multiple techniques.* When different anonymization techniques are simultaneously used, for example to satisfy different application requirements, an adversary with access to the different produced datasets may be able to infer private information. Further work is needed in understanding these risks and offering appropriate solutions.
- *Analysis and penetration testing*. The described privacy algorithms are relatively new and should be subjected to more rigorous security analysis. As with other security techniques, only continued analysis and penetration testing over time will provide a good understanding of the exact level of protection they offer.

Cross References

► Multiple Target Tracking

Recommended Reading

1. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the ACM SIGMOD Conference on Management of Data, May 2000, pp. 439–450. ACM Press, Dallas, Texas (2000)
2. Beresford, A., Stajano, F.: Mix zones: User privacy in location-aware services. In: IEEE Workshop on Pervasive Computing and Communication Security (PerSec), pp. 127–131. IEEE Computer Society, Washington, DC, USA (2004)
3. Blumberg, A.J., Keeler, L., Shelat, A.: Automated traffic enforcement which respects driver privacy. In: Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, September 2005, pp. 650–655. IEEE Computer Society, Washington, DC, USA (2005)
4. Brandeis, L., Warren, S.: The right to privacy. Harvard Law Rev. **4**(5), 193–220 (1890)
5. Chaum, D.: Security without identification: Transaction systems to make big brother obsolete. Commun. ACM **28**(10), 1030–1044 (1985)
6. Choi, J.Y., Jakobsson, M., Wetzel, S.: Balancing auditability and privacy in vehicular networks. In: Q2SWinet '05: Proceedings of the 1st ACM international workshop on Quality of service & security in wireless and mobile networks, Montreal, Quebec, Canada, 2005, pp. 79–87. ACM Press, New York, NY, USA (2005)
7. Civilis, A., Pakalnis, S.: Techniques for efficient road-network-based tracking of moving objects. IEEE Trans. Knowl. Data Eng. **17**(5), 698–712 (2005) Senior Member-Christian S. Jensen
8. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the First International Conference on Mobile Systems, Applications, and Services, San Francisco, CA, pp. 31–42. ACM Press, New York, NY, USA (2003)
9. Hoh, B., Gruteser, M.: Protecting location privacy through path confusion. In: Proceedings of IEEE/Create-Net SecureComm 2005, Athens, Greece, September 2005, pp. 194–205. IEEE Computer Society, Washington, DC, USA (2005)
10. Hoh, B., Gruteser, M., Xiong, H., Alrabady, A.: Enhancing privacy preservation of anonymous location sampling techniques in traffic monitoring systems. In: Proceedings (Poster Session) of IEEE/Create-Net SecureComm 2006, Baltimore, Maryland August 2006, pp. 1–3. IEEE Computer Society, Washington, DC, USA (2006)
11. Hoh, B., Gruteser, M., Xiong, H., Alrabady, A.: Enhancing security and privacy in traffic-monitoring systems. IEEE Pervasive Comput. **5**(4), 38–46 (2006)
12. Hoh, B., Gruteser, M., Xiong, H., Alrabady, A.: Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking. In: CCS '07: Proceedings of the 14th ACM conference on Computer and communications security, Alexandria, Virginia, USA. ACM Press, New York, NY, USA (2007, in press)
13. Reid, D.: An algorithm for tracking multiple targets. IEEE Trans. Autom. Control **24**(6), 843–854 (1979)
14. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (Technical Report SRI-CSL-98-04). Computer Science Laboratory, SRI International (1998)
15. Li, M., Sampigethaya, K., Huang, L., Poovendran, R.: Swing & swap: user-centric approaches towards maximizing location privacy. In: WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society, Alexandria, Virginia, USA, pp. 19–28. ACM Press, New York, NY, USA (2006). doi: 10.1145/1179601.1179605

Privacy Preserving in Location Based Services

► Privacy Preservation of GPS Traces

Privacy Threats in Location-Based Services

CLAUDIO BETTINI[1], SERGIO MASCETTI[1], X. SEAN WANG[2]
[1] Department of Informatics and Communication (DICo), University of Milan, Milan, Italy
[2] Department of Computer Science, University of Vermont, Burlington, VT, USA

Synonyms

Anonymity in location-based services; Location trusted server; Location server; Geopriv group, IETF; Identity aware LBS; Identity unaware LBS; Access control; Gnerelization; K-anonymity

Definition

Location-based services (LBS) are those services that, based on the user's current position, can provide location-aware information. Typical examples are map and navigation services, services that provide information on close-by public resources (e. g., gas stations, bus stops, pharmacies, and ATM machines), services that provide localized news (e. g., weather forecasts, road constructions, etc.), emergency services (911, 118, etc.) as well as more personalized services like proximity marketing or friend-finder.
Private information refers to the information a user does not wish to be released associated with her identity. This includes political or religious orientation, health information, financial assets, or closeness to specific individuals or organizations. LBS services play a role in this context because both identity and private information can be directly or indirectly released through a single or a sequence of LBS requests. LBS requests can reveal, for example, a) information on the specific location of individuals at specific times, b) movement patterns (specific routes at specific times and their frequency), c) requests for sensitive services (closest temple for a specific religious worship), or d) personal points of interest (frequent visits to specific shops, clubs, or institutions).

A privacy threat occurs whenever the information contained in one or more requests issued by a given user can be used, possibly associated with external information, to associate the user identity with the private information. The study of privacy threats and protection techniques in LBS is a subtle and challenging research topic.

Historical Background

Most of the approaches proposed in the literature to protect LBS privacy consider scenarios that can be easily mapped to the one depicted in Fig. 1.

Three entities are involved in the scenario:

- The **User** invokes or subscribes to location-based remote services that are going to be provided to her mobile device.
- The **Location Trusted Server (LTS)** is supposed to have access to the precise location data of a large group of users, act as a proxy for all LBS requests from these users and to enforce privacy policies for those requests.
- The **Service Provider (SP)** fulfills user requests and communicates with the user through the LTS.

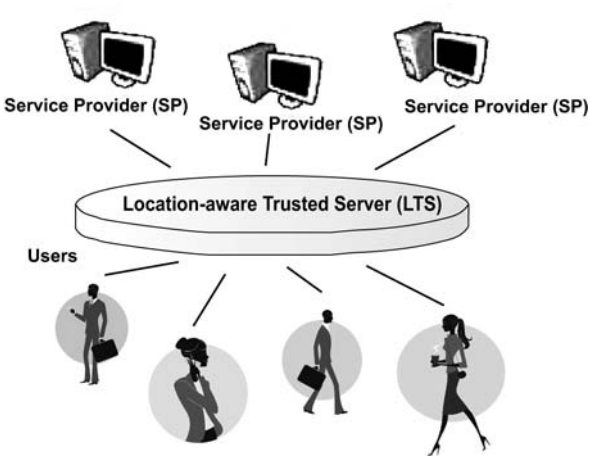
As pointed out above, this can be considered the reference scenario for many of the existing approaches to the privacy problem. Considering the model proposed by the IETF geopriv working group [6], the so-called *Rule Holder* and *Location Server* can be mapped as a single entity to the Location Trusted Server in the reference scenario; the *Location Generator* and *Location Recipient* can be mapped to the User device/infrastructure and Service Provider, respectively. The existence of an intermediate entity between the user and the service provider, possibly organized in multiple modules, is assumed also in [4,7,10,11,13,14,16].

Each user's request contains parameters concerning the identity of the user, the current time and position, and the requested service. While the specific request message may have an articulated structure depending on the specific parameters, three logical components can be identified:

$$r = \langle IDdata, STdata, SSdata \rangle$$

IDdata contains information on the user identity, *STdata* contains the requested spatio-temporal information, and *SSdata* specifies the service provider and the specific service parameters.

Any request in the above logical format may be the potential source used by attackers to violate user privacy. Requests, once forwarded by the LTS, may be acquired by potential attackers in different ways: they may be stolen from SP storage, voluntarily published by the trusted parties, or they may be acquired by eavesdropping on the communication lines. On the contrary, the communication between the user and the LTS is considered as trusted and the data stored at the LTS is not considered accessible by the attacker. Any entity that has access to the data of the SP or that can intercept the data in the communication channel between the LTS and the SP is a potential attacker. Hence, without loss of generality, in the following sections, the SP is considered as the potential attacker.



Privacy Threats in Location-Based Services, Figure 1 The reference scenario

Scientific Fundamentals

A subtle problem in LBS privacy is the identification of the possible threats. Section A provides a categorization of the possible privacy threats; first considering isolated requests, then sequences of requests issued by the same user. In Section B, the various techniques proposed in the literature are reported and their goals are specified with respect to the identified privacy threat.

A Privacy Threats in LBS

Considering the literature that has appeared on this topic, this section characterizes the typical scenarios of LBS requests and the possible privacy threats. In the following, there is a distinction between services that need to be aware of the user's identity (*identity aware LBS*) and services that can work with a pseudo-id or in an anonymous way (*identity unaware LBS*).

A.1 Privacy Threats in a Static Scenario The specific threats involved in the submission of each LBS request are considered here.

s.1 Sensitive association: IDdata and STdata. When a user submits an LBS request, an attacker could associate the user's identity with her spatio-temporal

location. There are some situations in which this association is considered sensitive by the user. Examples include all the spatio-temporal locations that can reveal a user's private information such as health status, political affiliations or religious beliefs.

s.1a Identity aware LBS In this case an attacker can always associate the identity of the user with her spatio-temporal location.
Consider a service for localized news that collects user identities in order to charge the cost of the service. If a user submits the service request while attending a political demonstration, an attacker could associate user identity with her political affiliation.

s.1b Identity unaware LBS Though user identity is not explicitly transmitted to the SP, an attacker can infer it from SSdata and STdata. The problem of discovering the identity of a user from a combination of values that refers to her has been extensively investigated in the area of database systems [12,15].
Consider an LBS that provides driving directions based on current traffic conditions. The service is provided by a business for its employees usually working outside the business establishment. The service is anonymous since the business cannot monitor the movements of its employees due to legal restrictions. However, if a user performs a request from his suburban house, then an attacker can compute the address from the STdata and associate it with the identity. This information can be used by the business to understand that the employee is not where she is supposed to be at that moment.

s.2 Sensitive association: IDdata and SSdata When a user submits an LBS request, an attacker could associate the user's identity with the specific service and service parameters invoked by the request. This data may be itself private information, or may be used to obtain private information.

s.2a Identity aware case This is the case in which the identity of the issuer is explicitly contained in the request. This case does not involve handling of spatio-temporal information and therefore is not peculiar to LBS related privacy threats.

s.2b Identity unaware case As in s.1b, the attacker is not aware of the identity of the user but can infer it from SSdata or STdata. Differently from s.1b, in this case a user's private information is derived from SSdata.
Consider a location based friend-finder service that is designed to introduce a user to the people that are close by and that have similar interests. Anonymous users send requests providing, as SSdata, some personal data (gender, age, first name) and an interests profile (sport, music, etc.). Suppose the user sends a request from her apartment. From the STdata, the attacker can discover the address from which the request is sent. Then, from a voter list (or other publicly available source), the attacker can acquire a list of people living at that address and obtain some information about them, such as age, gender and first name. These values can be compared with the ones provided by the user to possibly discover an identity that is eventually associated with the user's interests.

Note that a special case is that the STdata and/or SSdata fields are empty. If both fields are empty, it could still be a source for privacy violation, e. g., in threat s.1a, in which the privacy information could simply be the fact that the user has made a request. This also applies to the threats in the dynamic scenarios detailed in the next subsection. Note, however, that such empty-field requests can be seen as having special implicit (or default) values for the requests themselves. For example, a request without any SSdata to a toll booth could implicitly contain SSdata "paying for toll". Empty STdata could implicitly mean that the location and time are irrelevant to the request and therefore can be taken similarly as "any possible location and any possible time".

A.2 Privacy Threats in a Dynamic Scenario It is very common that a user submits several requests to the same LBS. If the requests are not related, each of them can be considered individually as in the static scenario. On the other hand, if it is possible for an attacker to link the requests to the same user, new privacy threats are possible. Several techniques exist to link different requests to the same user. The most intuitive one is the observation of the same identity or pseudo-id. Since the ability of linking requests is not a threat in itself, these specific techniques are not herein addressed.
A *request trace* is a set of requests that the attacker can correctly associate to a single user. Analogously, a *STdata trace* (and *SSdata trace*) is the set of STdata (and SSdata) that is contained in a request trace.
In a dynamic context, a user can consider as sensitive the association of her identity with a trace of STdata or SSdata. In this case, the LTS has to guarantee that the user identity is not disclosed in the same request trace.

d.1 Sensitive association: IDdata with STdata or SSdata. Note that these are the same associations as in s.1 and s.2; however, in a dynamic scenario new threats are possible that reveal these associations.

There are situations in which a location, such as an office or a house, may identify a small group of people. However, a single request from that location may not be sufficient to identify the sender. For example a user can submit an LBS request from a colleague's office or from a house where she is not living in. However, if a user submits LBS requests several times from that location, an attacker can trace the requests and eventually identify the user.

For instance, an attacker can observe a user's movements that repeat frequently. Using empirical assumption about users habits (like "users go from home to work in the morning and back in the afternoon") some information can be acquired about the user.

Consider a service that provides driving directions based on current traffic conditions. A professor uses the service every day to go from home to the university and back. The attacker sees that the pattern going from the location of the house to the location of university and back is frequent (almost every working day) hence it can infer that the two locations correspond to house and work. A cross check of the list of university faculty and the list of people living in the house leads the attacker to obtain the identity of the professor.

d.2 Sensitive association: IDdata and STdata trace. There are situations in which STdata of a single request can reveal personal information (e. g., threat s.1). However, in general, a trace of STdata or SSdata can provide the attacker with more reliable information.

An attacker that observes through a STdata trace that a user goes to a Church every Sunday can deduce the user's religious beliefs with good reliability.

d.3 Sensitive association: IDdata and SSdata trace. In many cases, the private information that an attacker can obtain from SSdata trace is equivalent to the union of the pieces of information that the attacker would obtain from each single SSdata. However, in general, there is some private information that can only be inferred from a trace of SSdata and not from a single one.

Consider a disease that obliges a user to frequently call a medical emergency service. In each request the user specifies the symptoms. While a single request could not reveal the disease, if the attacker observes the SSdata trace, he could discover it.

d.4 Disclosure of visited locations. If an attacker knows a user's STdata trace, it could infer other positions the user visited even if no request was performed from those locations.

Consider a service for car accident monitoring. Each user is identified by the SP with a pseudo-id. Suppose that a user is traveling on a highway and frequently communicates her position to the SP. On the highway, there are cameras that recognize car plates (e. g., to charge the road toll). The user is aware that revealing her car plate number together with the service pseudo-id can lead an attacker to associate her identity with the trace (the user suspects that the attacker can also access the data from the cameras). Hence, the user temporarily suspends the service. However, the attacker knows the user's locations before and after the cameras and therefore can infer that the user traveled through the area where cameras are positioned. Considering the average speed, the attacker can also estimate when the user was there and hence can associate the identity with the trace.

B Privacy Preserving Techniques

In this section the different techniques that have been proposed to address the privacy threats are briefly illustrated.

B.1 Access Control Advanced access control models can be used in the context of LBS services to specify and enforce privacy policy rules. The rules can define, for example, the type of data that each service provider can access, the resolution of that data, and possibly other constraints. With respect to our reference model, policies can be defined by users as well as by service providers and can be enforced by the LTS. Among the efforts in this direction, in [16] a push-based LBS scenario is considered; users can define authorizations that not only select which service providers can access location/profile information, but can also constrain the area and time in which they can send their offers to the users. The LTS is in charge of enforcing the authorizations. Among other efforts, the IETF Geopriv working group is proposing a format for expressing privacy preferences for location information [6]. With respect to the threats identified in the previous section, access control is an important component of a privacy preserving solution for all the threats. However, the best results in addressing the threats would probably be obtained by coupling access control with the anonymization techniques discussed below.

B.2 Temporal and Spatial Generalization The threats s.1b and s.2b illustrated in Section A.1 have many analogies with the problem of guaranteeing anonymity of personal data extracted from a relational database (see,

e. g., [15]). Typical solutions involve either the de-identification of data, essentially avoiding the presence of values that may directly or indirectly identify the user, the obfuscation of sensitive data, or the separation of identifying values from sensitive data. The first two solutions are usually based on the generalization or suppression of attribute values. Despite there are specific issues that distinguish the location-based problem from the analogous one in the relational database scenario, similar techniques can be applied. Indeed, the dynamic change of spatio-temporal resolution is an obfuscation technique based on generalization.

The idea of adapting spatio-temporal resolution to provide a form of location k-anonymity can be found in [8]. In the field of relational databases, a tuple is said to be k-anonymous if, considering the values of the attributes that could lead to re-identification, it is indistinguishable from other $k-1$ tuples. Analogously, an LBS request is considered k-anonymous if in the same area and temporal interval of the request there are k users that could have submitted that request. The privacy preserving technique consists of enlarging the location area and time interval of a request in order to include $k-1$ other potential users. This work is extended in [7,11,13]. The proposals in [7,13] support the use of a different value of k for different users. In [7], a slightly different notion of k-anonymity is used: the authors require the other $k-1$ users to have actually sent a request. In [13], each user can also specify the parameter $A_{\min}$ that indicates the minimum size of the area that the LTS should forward to the SP. In [11], a possible problem with the generalization proposed in [8] is pointed out and a solution is presented. The notion of k-anonymity in LBS has been more formally defined in [4], where, as in [8], the only requirement is the presence in the same spatio-temporal context of $k-1$ *potential* senders, which is a much weaker requirement.

The application of the techniques proposed in [7,8,11,13] avoids the association by an attacker of the identity of a request sender with a group of identities smaller than k. Hence, the addressed threat is s.2b and indirectly s.1b. The proposed solution partially solves the problem. Indeed, by applying this privacy preserving technique, it is ensured that STdata cannot be used to infer the identity of the request sender. However, the case in which a combination of values of STdata and SSdata is used is not considered.

The proposed solution ensures anonymity of the request sender even in the case in which the attacker knows the location and the identity of every person. If this very conservative assumption is adopted, it is always possible for an attacker to associate the identity of a user to their STdata (threat s.1) even if no requests are submitted. Therefore, in this case, the scope of the technique is limited to the prevention of threat s.2b.

B.3 Identification and Prevention of Critical Request Traces An important aspect in the dynamic scenario is how an attacker can identify a request trace and how a privacy preserving system can avoid it. Two cases have been considered.

In [1,2], LBS's that require a pseudo-id are considered. The proposed privacy preserving technique is based on the notion of *mix-zone* introduced by the authors and aims at avoiding the instance that an attacker traces the requests from the same user for a long period of time. The central idea is to change a user's pseudo-id each time the user enters a mix-zone. A mix-zone is analogous to a mix-node in communication systems [5] and can be intuitively described as a spatial area such that, if an individual crosses it, then it won't be possible to link his future positions (outside the area) with known positions (before entering the area). Here, "link" defines the association of different requests to a single trace.

The results can be applied in the dynamic scenario, but cannot be used to provide a complete solution to any of the threats described in Section A.2. Indeed, the technique aims at reducing the length of the request traces but does not evaluate if sensitive information is released. Nevertheless, reducing the request trace length is an important task that facilitates privacy protection in a dynamic context. Hence, this technique could be very useful as a part of a privacy preserving system (like in [4]).

A different approach to the issue of request traces is to consider LBS's that do not require pseudo-ids. This case is considered in [9], where the authors experimented to see if it is possible for an attacker to trace a user. A known algorithm for tracking multiple objects is applied to trace a small number of users whose locations are frequently collected. The authors concluded that it is practically possible for an attacker to obtain request traces even if pseudo-ids are not submitted to the SP. This paper does not propose a solution to preserving privacy but is a preliminary step in the definition of a technique that could be used by a privacy preserving system to evaluate if a user is possibly being traced by an attacker. In the absence of such a solution, a privacy preserving system should adopt a conservative approach assuming that a user can always be traced.

B.4 Techniques to Prevent Location Identification In [10], it is assumed the case in which each user specifies the *sensitive areas*, i. e., geographic positions that should never be associated to the presence of that specific user. The aim of the proposed privacy preserving system is to avoid that an attacker can understand that the user visited a sensitive area. The straightforward solution of suppressing all the requests from these areas is not effective since an attacker could infer that a user visited a sensitive area

only from her request trace outside the sensitive area. This situation is analogous to the one used for threat d.4.

The proposed solution is based on a partitioning of all the areas (sensitive or not) in zones such that each zone includes at least k sensitive areas. Then, each request is suspended until the user crosses a zone boundary. If the user has not visited a sensitive area, all the pending requests are submitted, otherwise they are suppressed.

To our knowledge, the proposed technique is the first one that addresses this kind of problem. However, it is not clear if it is an effective solution. First, it is debatable if it is appropriate to extend the concept of k-anonymity to sensitive areas. Indeed, if a user specifies some sensitive areas, she does not want her identity to be associated with any of them; On the contrary, the proposed solution allows an attacker to infer that a user visited a sensitive area even if it cannot say which one it is in a set of k. Secondly, it is not clear if it is acceptable to always postpone the submission of a request until a user changes a zone. Finally, the way in which the zones are constructed is critical. Indeed, it seems possible in some specific cases that an attacker could infer the exact sensitive area a user visits.

B.5 k-Anonymity Techniques in a Dynamic Scenario

The extension of location k-anonymity to the dynamic scenario has been investigated in [4]. The investigation presents a basic question. Is a trace k-anonymous if each request in the trace is k-anonymous according to the definition for the static scenario? The answer is negative; indeed, if m requests constitute the trace, the attacker may have available m sets, each one with at least k candidate individuals. However, since he knows that all the requests in the trace were made by the same individual, he can perform the intersection of the m sets, possibly obtaining less than k individuals.

The formal property needed to guarantee k-anonymity of a trace of LBS requests is called *historical k-anonymity*. Some preliminary definitions are necessary to formally define it. It is reasonable to assume that the LTS not only stores in its database the set of requests that are issued by each user, but also stores for each user the sequence of her location updates. This sequence is called *Personal History of Locations* (PHL). More formally, the PHL of user U is a sequence of 3D points $(\langle x_1, y_1, t_1\rangle, \ldots, \langle x_m, y_m, t_m\rangle)$, where $\langle x_i, y_i\rangle$, for $i = 1, \ldots, m$, represents the position of U (in two-dimensional space) at the time instant t_i.

Note that a location update may be received by the LTS even if the user did not make a request when being at that location. Hence, for each request r_i there must be an element in the PHL of the user issuing r_i, but the vice versa does not hold. This has an intuitive motivation in the fact that the anonymity set for a certain area and a certain time interval is the set of users who were in that area in that time interval and who could *potentially* make a request.

A PHL $(\langle x_1, y_1, t_1\rangle, \ldots, \langle x_m, y_m, t_m\rangle)$ is defined to be *LT-consistent* with a set of requests $r_1, \ldots, r_n$ issued to an SP if, for each request r_i, there exists an element $\langle x_j, y_j, t_j\rangle$in the PHL such that the area of r_i contains the location identified by the point x_j, y_j and the time interval of r_i contains the instant t_j.

Then, given the set R of all requests issued to a certain SP, a subset of requests $R' = \{r_1, \ldots, r_m\}$ issued by the same user U is said to satisfy *Historical k-Anonymity* if there exists $k-1$ PHLs $P_1, \ldots, P_{k-1}$ for $k-1$ users different from U, such that each $P_j, j = 1, \ldots, k-1$ is LT-consistent with R'.

In practice, it is clearly difficult to keep request traces k-anonymous, thus, techniques like mix-zone or frequent change of pseudo-id reducing the length of traces are indeed very important. In [4], a preliminary investigation is reported on the techniques that could be used to preserve historical k-anonymity.

Key Applications

LBS's have been extensively used both in military as well as in commercial applications. With the diffusion of location-aware devices and with the increasing precision of localization techniques, many new commercial applications based on LBS's are being deployed. These applications are not only targeted to business users, but also to generic users interested in LBS for their spare time activities.

LBS privacy preserving techniques can be implemented in a middleware layer to provide privacy protection for most of these applications. Clearly, the level of protection for a given application and for a given user category will be driven by different parameters and possibly with different techniques.

Moreover, the assumptions on the available external knowledge will also determine the most appropriate features that the privacy protection middleware should have. As presented in Section A, the functionality of such a middleware can be focused on protecting the association of the user identity with

- location or trace information,
- service access information,
- both of the above.

Future Directions

Different research directions can be identified in this field.

- **Definition of a formal framework.** Existing approaches would benefit from a rigorous formalization of problems and solutions; In particular, formal definitions of

privacy threat and *defense technique* should be provided. The formalization should also be sufficiently expressive, modeling different kinds of knowledge that may be available to the attacker. Indeed, existing approaches do not explicitly define the knowledge that may be available to the attacker; consequently, the proposed defense techniques are subject to critique based on counterexamples, assuming that there is specific knowledge available to the attacker.

- **Design of safe defense techniques.** Based on the formal characterization of a privacy problem either in the static or in the dynamic case, new defense techniques should be designed for that specific problem. Each defense technique should be safe with respect to the specific assumptions made about the knowledge that may be available to the attacker. Moreover, defense techniques should be optimized with respect to the global processing costs at the LTS.
- **Collection and/or generation of significant data for experiments**. The applicative and critical nature of this research field makes it crucial to verify the effectiveness of the proposed defense techniques with real-world data. However, real-word data should include the movement traces of a large number of users for a long period of time and a significant collection of sensitive service requests that these users have issued. Obtaining this information or generating it with a sufficiently accurate model is a challenging task.

Acknowledgements

This work was partially supported by Italian MiUR InterLink project N.II04C0EC1D, and the US NSF grants IIS-0430402 and IIS-0430165.

Cross References

► Indexing, Hilbert R-tree, Spatial Indexing, Multimedia Indexing

Recommended Reading

1. Beresford, A.R., Stajano, F.: Location Privacy in Pervasive Computing. IEEE Pervasive Computing **2**(1):46–55 January–March (2003)
2. Beresford, A.R., Stajano, F.: Mix zones: User privacy in location-aware services. In: PerCom Workshops, pp. 127–131. IEEE Computer Society (2004)
3. Bettini, C., Mascetti, S., Wang, X.S., Jajodia, S.: Anonymity in Location-based Services: Towards a General Framework. In: Proc. of the 8th International Conference on Mobile Data Management (MDM). IEEE Computer Society (2007)
4. Bettini, C., Wang, X.S., Jajodia, S.: Protecting privacy against location-based personal identification. In: Jonker, W., Petkovic, M. (eds.) Secure Data Management, Lecture Notes in Computer Science, vol. 3674, pp. 185–199. Springer (2005)
5. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM **24**(2):84–88 (1981)
6. Cuellar, J., Morris, J., Mulligan, D., Peterson, J., Polk, J.: Geopriv requirements. RFC 3693, IETF (2004)
7. Gedik, B., Liu, L.: Location privacy in mobile systems: A personalized anonymization model. In: ICDCS, pp. 620–629. IEEE Computer Society (2005)
8. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: Proc. of the First International Conference on Mobile Systems, Applications, and Services (MobiSys), USENIX (2003)
9. Gruteser, M., Hoh, B.: On the anonymity of periodic location samples. In: Hutter, D., Ullmann, M. (eds.) SPC, Lecture Notes in Computer Science, vol. 3450, pp. 179–192. Springer (2005)
10. Gruteser, M., Liu, X.: Protecting privacy in continuous location-tracking applications. IEEE Security & Privacy **2**(2):28–34 (2004)
11. Kalnis, P., Ghinta, G., Mouratidis, K., Papadias, D.: Preserving anonymity in location based services. Technical report TRB6/06, University of Singapore (2006)
12. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 49–60. ACM Press (2005)
13. Mokbel, M.F., Chow, C., Aref, W.G.: The new casper: query processing for location services without compromising privacy. In: Proc. of the 32nd International Conference on Very Large Data Bases (VLDB), ACM (2006)
14. Myles, G., Friday, A., Davies, N.: Preserving Privacy in Environments with Location-Based Applications. IEEE Pervasive Computing **2**(1):56–64 January–March (2003)
15. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering **13**(6):1010–1027 (2001)
16. Youssef, M., Atluri, V., Adam, N.R.: Preserving mobile customer privacy: an access control system for moving objects and customer profiles. In: MDM '05: Proceedings of the 6th international conference on Mobile Data Management, pp. 67–76, New York, NY, USA, ACM Press (2005)

Private Visual Thinking; Reexpression

► Exploratory Visualization

Probabilistic Map Algebra

► Bayesian Network Integration with GIS

Probability Networks

► Bayesian Network Integration with GIS

Probability Theory

► Objects with Broad Boundaries
► Uncertain Environmental Variables in GIS

Probability Treshold Indexing

► Spatial Data, Indexing Techniques

Probe Vehicle Data

► Floating Car Data

Problem of Seven Bridges of Konigsberg

► Graph Theory, Konigsberg Problem

Process

► Geographic Dynamics, Visualization And Modeling
► Temporal GIS and Applications

Process Model

► Hierarchical Spatial Models

Processes and Events

ANTONY GALTON
School of Engineering, Computer Science,
and Mathematics, University of Exeter, Exeter, UK

Synonyms

Activities and occurrences

Definition

The terms 'process' and 'event' both refer to things that happen or go on in time, and are invariably associated with some kind of change. The terms are often used loosely, sometimes interchangeably, and there is little consensus on how they should be defined or distinguished from one another. The following definitions are suggested to bring the terms into line with their main everyday uses while providing a useful basis for technical discussions.

Process: An ongoing dynamic situation involving the activity of one or more material things or portions of matter. Key properties of processes are that (1) they are conceptualized as homogeneous, i. e., if a process is going on over some interval of time then it is also going on over all the subintervals of that interval (at least down to some minimal length associated with the inherent 'grain-size' of the process), and (2) they are open-ended, i. e., if a process is going on at a particular time then in principle it can continue going on into the future—thus processes are not intrinsically bounded by an end-point or completion.

Event: An individual occurrence or episode singled out from the ongoing processual flux, with a definite beginning and end (which coincide in the case of instantaneous events). Key properties of events are exactly opposed to those of processes: (1) they are not homogeneous, since if an event occurs over an interval then it does not occur over any of the subintervals (so the event's occurrence is a global property of the interval, not a local property of its subintervals), and (2) events are not open-ended—once an event has occurred it cannot go on occurring (although another event of the same type may begin).

Historical Background

Questions about the general classification of processes, events, and related categories have been pursued in three main areas: philosophy, linguistics, and artificial intelligence. The philosophical contribution is largely in the areas of metaphysics, ontology, and the theory of action. In linguistics, the focus is on classifying verbs and verb-phrases in order to explain the different ways in which they are used, and in particular to account for their interaction with temporal elements such as tense and aspect, temporal prepositions, and so on. In artificial intelligence, the emphasis is on representing knowledge about processes and events in such a way as to facilitate reasoning about them for purposes such as planning, prediction, and explanation.

On the philosophical background, a useful point of entry is provided by [2], which collects together a wide range of papers on events. In linguistics, [6,7] provide useful background on the classification of expressions denoting processes and events. In artificial intelligence, two classic papers are [1,5]. A useful collection of papers which includes work from all three background areas is [4].

In GI Science, while the necessity for handling temporal phenomena has been acknowledged for some time now, progress has been hampered by the lack of principled ways of describing these phenomena; in particular, the failure to distinguish adequately between processes and events has led to a certain amount of confusion in the literature.

As Worboys [10] says, 'One person's process is another's event, and vice versa'. It is important not to let the terminological variation detract from the real merit of much of the work that has dealt with dynamic phenomena in GIS, but it remains true that progress would be expedited if a common usage of the terms could be agreed.

Scientific Fundamentals

Any application which handles both processes and events should also handle the relationships between them. Some of these relationships are very general, belonging at a high level of abstraction; such relationships can be systematized into an *algebra* of processes and events. The operations of such an algebra include ways of deriving events from processes and ways of deriving processes from events. An example of the former would be an episode of erosion leading to the ultimate separation of a promontory from the mainland; an example of the latter would be flow of traffic past a junction, which is a process consisting of many individual events corresponding to the passage of individual cars past the junction.

The open-ended, homogeneous nature of processes means that processes should be regarded as existing from moment to moment, in contrast to an event which spans a whole interval and is not present as a whole at individual moments within the interval. It makes sense to say that a process is in operation at a single time of observation (even though certain *methods* of observation, such as the capturing of a 'snapshot' image, may fail to reveal the processes in operation), whereas events are normally identified retrospectively by synthesizing a sequence of states over an interval or comparing some initial state of affairs with an eventual outcome. Thus a process like the flow of a river can be observed directly, and its properties (e. g., its speed) measured at the moment of observation. A corollary of this is that processes can change: the properties of a process measured at a later time may differ from those measured earlier. This in turn allows the definition of second-order processes such as the increase in flow of the river.

For events, by contrast, a different, historical perspective is adopted. An event is a discrete unity[1] that is located at a particular time, either a point or an interval, so it does not make sense to ascribe change to it: rather, it just happens and then so to speak 'sits there' in the historical record. (In cases where it seems as though an event is changing, it is invariably the constituent process of the event, rather than the event itself, to which change is more properly ascribed—e. g., to say that a volcanic eruption became more violent is to say that the erupting process changed in this way; the eruption as an event is a unitary whole which subsumes the profile of variable violence within its temporal 'shape' but cannot meaningfully be said to undergo change itself.)

Note however the word process is often used to refer to what are in fact *procedures*, e. g., the process of refueling an aircraft. This is something that follows a prescribed pattern, involving various events and processes; it is typically something for which explicit instructions could be given. Each individual instance of a procedure being carried out is an event, not a process in the sense used in this article. Although the word 'procedure' suggests something involving human agency, similar 'structured events', proceeding according to a fixed pattern, occur in nature, although perhaps more typically in biological than geographical contexts, e. g., cell division.

With both processes and events, it is necessary to distinguish generic *types* from specific individual *instances* of those types. This is particularly clear in the case of events: there is an obvious difference between the notion of a volcanic eruption as a generic type of event that may occur at many different places and times, and specific instances of this type such as the eruption of Vesuvius in 79 A.D. or that of Mount St Helens on May 18th 1980. This distinction is sometimes obscured by normal ways of speaking: 'the same thing happened again' does not refer to the same individual event but to another individual of the same *type*. With processes, similarly, longshore drift as a generic process-type can be distinguished from the current realization of that process along a particular stretch of coastline.

Processes can be broadly classified as steady-state, cyclic, irregular, or progressive, depending on the profile of their temporal development.

A steady-state process maintains an existing state of affairs: it can operate over arbitrarily long periods while producing no net change. An example is the flow of water through a lake, where the inflow is exactly balanced by the outflow, resulting in the water-level remaining constant; although there is undeniably a process in operation here, to an outside observer it looks static. Even where movement is discernible, a process can still be classed as steady-state if there is no overall change, at a certain level of description: the flow of traffic along a stretch of motorway, for example, can be regarded as steady-state so long as the speed and density of the traffic remain more or less constant.

In a cyclic process, there is a regular periodic variation in the associated state of affairs. An example is the rise and fall of the tides along a shoreline, endlessly repeating the cycle of high tide followed by low tide followed by high tide again. The periodicity does not have to take a simple sinusoidal form: several cycles can be superposed to produce more complex profiles of variation. This is true of the

[1] The quality or state of not being multiple, i. e. oneness.

tidal process, since the half-daily alternation of high and low is accompanied by a monthly alternation of spring and neap as well as longer-term alternations resulting from the complex interaction of the relative motions of the earth, moon, and sun.

Sufficiently complex superpositions of periodicities can result in what appears to be a completely irregular process profile, and of course it becomes a matter for scientific investigation as to whether an apparently irregular process is, in fact, the product of some complex combination of regular cycles. Many meteorological phenomena are of this kind, which is what makes them hard to predict.

The final category of processes are those which may be described as progressive, in which the operation of the process results in a cumulative change in some state of affairs. The process of urbanization by which some built-up area steadily encroaches on what was open country is progressive in this sense: the process does result in net change, and the longer the process operates for, the greater the magnitude of the change.

How a process is classified can depend critically on the temporal 'grain size', or granularity, under which it is described. Many apparently steady-state processes appear on closer examination to be cyclic or irregular. If the minimal temporal interval under consideration is greater than half a day, the high tide/low tide periodicity becomes indiscernible, but the longer term spring tide/neap tide periodicity may remain. Again, a steady-state traffic flow, when described at a finer temporal granularity, consists of the passage of an endless succession of individual vehicles, and this may be viewed from a certain point of view as a cyclic or irregular process.

Similarly a process which appears merely cyclic under a fine temporal granularity can be revealed as progressive when the focus is shifted to longer time periods. The flow of traffic along a particular stretch of road is cyclic to the extent that it undergoes regular and predictable daily and seasonal variation; superimposed on this there will be many smaller-scale irregularities, but more significantly, over a longer period of many years there may be a progressive increase in traffic levels. This might be understood as one process which has steady-state, cyclic, irregular and progressive aspects depending on the temporal 'focus' under which it is described, or as the superposition of a number of separate processes operating over different time scales. The former view is more concrete in the sense that the process simply is the flow of traffic that is observed; under the latter view the processes become more abstract and only accessible by inference from the observed flow.

None of the above applies to events. It is true that an event can be described as periodic, but in this case what is meant is that there is a process consisting of the regular repetition of some event type: each individual repetition is one event, and as such is not itself periodic.

Events may be classified as *punctual* (i. e., point-like) or *durative* (i. e., extended or interval-like) depending on whether or not they have appreciable duration. This too is a matter of granularity: on a scale of hours or days, a volcanic eruption may be an extended affair, but from a perspective spanning many years it may be effectively instantaneous.

Events may also be classified in terms of whether they are characterized *internally* in terms of their constitution, either as a 'block' of some process or as a composite of two or more subevents, or *externally* in terms of their effects, e. g., an event may be defined in terms of a transition between specified start and end states. A volcanic eruption is internally characterized as an episode which consists of the eruption process operating at a particular location over some particular interval of time. The separation of Great Britain from the European mainland is externally characterized as the transition from a state of affairs in which Britain and Europe formed part of a single land mass to one in which they were separated by the sea. An event which is externally characterized will usually in fact come about as a result of the operation of some process; but the external characterization does not explicitly refer to this. It is also possible to describe an event using both internal and external characteristics: the event reported in the statement 'John walked to the station' is characterized externally by the fact that it is a transition from a state of affairs in which John is not at the station to a state of affairs in which he is; and it is characterized internally by the fact that it comes about as a result of John's spending a while walking.

Key Applications

Processes, by nature, involve some form of change; in many cases, the change in question is motion, and it is usually continuous (or, if discrete, is only revealed to be so at a finer temporal granularity than that at which it is being represented as a process). The application areas in which the notion of process is likely to be invoked are therefore those in which there is change that is effectively continuous. This has implications for the kinds of technologies that can be used for modeling or otherwise representing processes.

The simplest way of handling time and change in the context of GIS is by means of a sequence of static "snapshots" representing the world at times $t_1, t_2, t_3, \ldots$. Assuming that individual snapshots contain no information about processes that are in operation at the times they represent, the

existence of processes can only be inferred by tracking the changes between successive snapshots. In effect this presents a sequence of events of the form 'between t_i and t_j the world changed from this state to that state'. These events by nature form a discrete series, which does not do justice to the continuous nature of processes. The processes present in the world (which give rise to the events, as it were forming their temporal substance), are not visible in such a model.

To represent processes explicitly, therefore, some additional technology is needed. One possibility is to include processes as first-class objects within individual snapshots: in effect, to treat a process as an individual entity which may be in operation from one snapshot to the next, itself perhaps undergoing change. An alternative approach is to move beyond the snapshot model and present the course of history directly, with all entities—processes, events, and ordinary objects—modeled as four-dimensional 'chunks' of reality. One can then obtain different views of the world by segmenting these chunks in different ways; for example, a conventional snapshot is now a 'slice' of the four-dimensional world orthogonal to the time axis, whereas the history of an object encapsulates the changes it undergoes in a 'life-line' that is extended along the time dimension. As yet there is no clear consensus as to how best to handle events and processes in the context of GIS, and it is therefore not possible to point to an established way of proceeding.

Since processes and events are ubiquitous in the world, almost any GIScience application may need to take them into account. The following paragraphs present a sample of areas where this need has been felt particularly keenly, and which have therefore figured prominently in recent discussions about how the temporal dimension should be integrated into GIS.

Weather systems. Yuan [11] distinguishes a precipitation event (i. e., 'the occurrence of precipitation in the study area') from its constituent precipitation processes (which describe 'how it rains'). This is broadly in line with the way in which the relationship between processes and events was characterized above. Yuan describes algorithms for assembling precipitation processes and events from raw precipitation data and for computing information about their behavior and interaction. This approach is advocated to provide enhanced support for 'complex spatio-temporal queries on the behavior and relationships of events and processes'.

Coastal Geomorphology. The sea coast is an excellent place to see both processes and events in action. The breaking of a particular individual wave is an event, but the regular succession of waves breaking along the shoreline is a process. On closer analysis, individual wave-breaks can be seen to be made up out of various kinds of processes, and if different types of wave are distinguished according to how they break (e. g., 'spilling' and 'plunging') then this explicitly invokes different constituent process types. In the longer term, erosion and deposition phenomena constitute processes which, in aggregate, may result in discrete events such as the separation of a headland from the mainland or the formation of a spit. Methods of representing such phenomena in an information system are described by [9].

Oceanography. Often the focus of interest is the profile of a process: its increases and decreases, its peaks and troughs, or periods of stability. The general picture assumed here is as follows: the raw data, as delivered by various sensors, takes the form of sequences of quantitative values of some time-varying observable. An example would be quantities such as the speed and direction of winds and current, wave height, air temperature and so on that might be measured by a buoy tethered to the ocean floor at a particular location. Ongoing changes in any of these quantities are processes. Salient demarcated episodes in the evolution of any of these process are events.

In the GoMOOS (Gulf of Maine Ocean Observing System) project (see http://www.gomoos.org/), a number of buoys at different locations across the Gulf of Maine are making regular observations of key oceanographic variables; the problem is then how to derive from these observations an understanding of the large-scale oceanic processes operating in the gulf. To facilitate such understanding it is essential to have an appropriate ontology for classifying processes and events, refined by careful consideration of the spatial dimension.

Traffic systems. Worboys [10] uses algebraic methods to specify traffic flow at a four-way stop. This is, essentially, an open-ended process built up from a sequence of atomic events. The atomic events are of two kinds: a vehicle arriving at the intersection, and a vehicle moving forward across the intersection. The specification shows how the ongoing process can be generated from an appropriate sequencing of such events. This describes the normal pattern of events, but in addition the modeling of traffic systems should take into account events such as traffic jams and accidents which arise from the particular conditions of the traffic flow process.

Traffic flow can be regarded as a process which exists at all points on the network at every time. It may be characterized by various attributes, of which the most important are

- speed (measured in units of distance/time)
- density (measured in units of vehicles/distance)
- rate of flow (measured in units of vehicles/time).

These attributes vary both spatially across the network and temporally. Each of the attributes has a value at each position and time, the rate of flow being the product of speed and density.

One thing which makes traffic flow particularly interesting is that it may be viewed from two quite different kinds of perspective, both of which will be important in different contexts. On the one hand, there is the bystander's point of view, that is, the point of view of someone positioned at a particular point in the network and watching the traffic flow past. This is the point of view that is relevant to a pedestrian trying to cross the road, or to a speed camera in a fixed position beside the road. It is also the point of view of someone living next to the road and concerned about traffic noise, the safety of their children, etc. The process that is of concern here is the traffic flow *past a point*; What the bystander observes is the fluctuation in the attributes of this process.

A different point of view is that of the participant in the traffic flow itself, i. e., the drivers of the vehicles whose motion through the network constitutes the flow. The flow as viewed from a given vehicle can be regarded as a process in its own right, whose time-varying attributes may be quite different from those of the process viewed by the bystander. To illustrate this, consider a long stretch of road with a single intersection half-way along, controlled by traffic lights. Then someone stationed at a point close to the traffic lights will observe a cyclical variation in the flow along this road, but the driver of a vehicle traveling along the road may, depending on the state of the lights when he reaches the intersection, either observe a uniform flow, or a flow which comes to a halt on one occasion and then a little while later starts up again. It is important that what is seen by each of these observers is a perfectly good process; and clearly there are systematic relationships between the processes seen by stationary observers at different points in the network and the processes seen by moving observers participating in the flow.

Future Directions

Temporal GIS is still an emerging field, and as such has not yet arrived at an established consensus on the terminology and classification of temporal phenomena—including broad categories such as process and event. As a result, development of temporal capability in GIS has mostly proceeded in a piecemeal and *ad hoc* fashion, with no solid basis in underlying theory. On the other hand, such theorizing as exists has tended to be divorced from the practicalities of implementing useful temporal functionality in GIS. A key desideratum for the future is therefore to establish an appropriate theoretical basis for describing the temporal dimension that is firmly grounded in practicality.

One element in particular will need to be looked at: in geographical information science it is customary to distinguish between object-based approaches, which think of space as populated by a collection of discrete objects each with a unique identity and attributes, and field-based approaches, which think rather of variables taking different values at different locations, providing a coverage of space. The ontologies implicit in most formal representation languages are heavily biased towards the object-based view of the world, but in application areas such as oceanography, meteorology, and coastal geomorphology a field-based view may be more appropriate. For this reason, any adequate treatment of processes and events must cover not only the changes undergone by individual objects, but also the the phenomena that may be observed in continuously varying fields. Indeed, as the examples above show, it is precisely in this latter kind of example that a proper treatment of processes is most urgently needed.

Cross References

- ▶ Geographic Dynamics, Visualization And Modeling
- ▶ Movement Patterns in Spatio-temporal Data
- ▶ Patterns in Spatio-temporal Data
- ▶ Sequential Patterns, Spatio-temporal
- ▶ Temporal GIS and Applications

Recommended Reading

1. Allen, J.: Towards a general theory of action and time. Artif. Intell. **23**, 123–54 (1984)
2. Casati, R., Varzi, A. (eds.): Events. The International Research Library of Philosophy. Dartmouth, Aldershot, UK (1996)
3. Hornsby, K., Worboys, M. (eds.): Event-oriented Approaches in Geographic Information Science. Lawrence Erlbaum Associates. Special issue of: Spatial Cognition and Computation, vol. 4, number 1 (2004)
4. Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.): The Language of Time: A Reader. Oxford University Press, Oxford (2005)
5. McDermott, D.: A temporal logic for reasoning about processes and plans. Cogn. Sci. **6**, 101–55 (1982)
6. Moens, M., Steedman, M.: Temporal ontology and temporal reference. Comput. Linguist. **14**, 15–28 (1988)
7. A. P. D. Mourelatos. Events, processes, and states. In: Tedeschi, P., Zaenen, A. (eds.) Tense and Aspect, pp. 191–212. Academic Press, New York (1981)
8. Peuquet, D.J., Duan, N.: An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. Int. J. Geogr. Inf. Syst. **9**, 7–24 (1995)
9. Raper, J.: Multidimensional Geographic Information Science. Taylor and Francis, London and New York (2000)
10. Worboys, M.F.: Event-oriented approaches to geographic phenomena. Int. J. Geogr. Inf. Sci. **19**(1), 1–28 (2005)
11. Yuan, M.: Representing complex geographic phenomena in GIS. Cartogr. Geogr. Inf. Sci. **28**, 83–96 (2001)

Profile, Topological

► Spatial Data Transfer Standard (SDTS)

Profiles

► Spatial Data Transfer Standard (SDTS)

Programmable GIS Components

► MapWindow GIS

Progressive Approximate Aggregation

IOSIF LAZARIDIS, SHARAD MEHROTRA
Department of Computer Science,
University of California, Irvine, CA, USA

Definition

Aggregate queries are answered by computing a single scalar value over a set of relevant data objects, e. g., the average temperature over all sensors in a region of space. Often, the precise value is not needed because the user may not be interested in knowing the temperature to the last decimal point. Additionally, the mode of visualization of the answer, e. g., in a virtual navigation, may in itself impose restrictions by necessitating either a high frame rate or the color-coding of the answer at a specific resolution. Progressive approximate aggregate queries compute the answer progressively, coming up with an initial estimate and refining it until the time deadline (e. g., time to render the frame), or answer quality precision (e. g., ± 1°C) is reached. Thus, they are a flexible way of query processing since they make no assumptions about the time/accuracy specifications imposed by the user application and can accommodate a wide variety of such specifications in a unified way.

Main Text

On-line aggregation was proposed as an approximate aggregation mechanism. Data objects are read and the aggregate function is computed with them. Using statistical techniques, probabilistic confidence intervals on the answer are computed; the user can stop the query whenever he is satisfied with the answer quality.

Progressive approximate aggregate queries using Multi-Resolution Aggregate tree (MRA-tree) data structures are another way of progressively approximating the answer. They avoid calculating the aggregate over individual tuples since these aggregates are pre-computed and stored in nodes of the MRA-tree. Moreover, at each step they provide a point estimate of the aggregate value (e. g., 34C) and a deterministic interval of confidence (e. g., the interval [33, 36]) guaranteed to contain the exact answer. Progressive approximation techniques are a powerful aid for geospatial visualization systems incorporating large amounts of data or requiring fast query answering, allowing such systems to function efficiently and to produce high quality answers with provable bounds using the available system resources.

Cross References

► Aggregate Queries, Progressive Approximate
► Multi-Resolution Aggregate Tree

PROJ

► Open-Source GIS Libraries

Properties, Geometric

► Geography Markup Language (GML)

Properties, Simple

► Geography Markup Language (GML)

Property Register

► Cadastre

Protective Action Zone

► Emergency Evacuations, Transportation Networks

Provable Properties

► Geosensor Networks, Formal Foundations

Proximity Matrix

► Spatial Weights Matrix

PR-Quadtree

▶ Quadtree and Octree

PSS

▶ Geocollaboration

Public-Domain Software

▶ deegree Free Software
▶ PostGIS

Public Health

▶ Pandemics, Detection and Management

Public Health and Spatial Modeling

ANDREW B. LAWSON
Arnold School of Public Health,
University of South Carolina, Columbia, SC, USA

Synonyms

Mapping; Spatial models; Disease mapping

Definition

In this chapter the emphasis is on the analysis of health data which are geo-referenced. Often data of this kind arise in the analysis of clustering of disease, or in the estimation of excess or relative risk in small administrative regions. Another aspect of this analysis may be the need to assess the relation between the spatial location of cases of disease and covariates, often spatially distributed. This is often called ecological analysis, especially when a change of spatial scale is involved. Finally, the possibility of carrying out prospective surveillance of disease maps is considered.

Historical Background

Spatial modeling of Public Health data is a wide subject and encompasses a range of topics where the geographical or spatial distribution of disease is of importance. In what follows there will be an emphasis on disease analysis as opposed to other aspects of Public Health sciences (such as health services research, health promotion or education). Statistical methods employed in this area are also diverse in their range and besides basic exploratory and descriptive methodology common to many subject areas, there is a need to employ particular *spatial* statistical methods which are designed for such data. The basic characteristic of data encountered in this application area is its *discrete* nature, whether in the form of spatial locations of cases of disease, or counts of disease within defined geographical regions. Hence methods developed for continuous spatial processes, such as Kriging, are not directly applicable or only approximately valid.

Often geographical hypotheses of interest in Public Health focus on whether the residential address of cases of disease yields insight into etiology of the disease or, in a public health application, whether adverse environmental health hazards exist locally within a region (as exemplified by local increases in disease risk). The classic example of the earliest epidemiological study was that of Snow [33] who examined the cholera epidemic in London in relation to water supply pumps. Since the 1970s there has been a growth in interest in geographical analysis of disease and this growth intensified in the 1980–1990 period with the development of geographical information systems (GIS) and public awareness of local environmental risks. The importance of geographical analysis of disease can be easily demonstrated. For example, in a recent study of the relationship between malaria endemicity and diabetes in Sardinia a strong negative relationship was found [4,25], ch 9. This relation had a spatial expression and the geographical distribution of malaria was important in generating explanatory models for the relation. In public health practice it is of considerable importance to be able to assess whether localized areas which have larger than expected numbers of cases of disease are related to any underlying environmental cause. Here spatial evidence of a link between cases and a source is fundamental in the analysis. Evidence such as a decline in risk with distance from the *putative* source of hazard or elevation of risk in a preferred direction is important in this regard.

Scientific Fundamentals

There are four main areas where spatial statistical methods have been developed for Public Health data analysis: relative risk estimation (sometimes simply called Disease Mapping), Disease Clustering, Ecological Analysis, and Disease Map Surveillance. Before looking in detail at each of these areas, it is appropriate to consider some common themes or issues which arise in all areas of the subject. Before considering the study of the spatial distribution of disease, there are some fundamental epidemiological ideas that should be considered.

Relative Risk

Within any geographical area the local density of cases of disease can be studied. There is a desire to examine this as it gives information about local variations in disease. With

census tracts then the count of cases of a particular disease could be the data of interest. These crude counts of disease cannot be used on their own as the density of cases will be affected by the variation in the population of the area. This is true whether observing case addresses (the residential address of a disease case) or the aggregated count of cases within small areas.

Hence underlying the disease incidence is the variation in the population 'at risk' of the disease. This background population will vary in its composition (age, gender, susceptibility groups) and in its density with spatial location. Hence this variation should be accounted for in any analysis of the disease occurrence. Clearly, if areas of high susceptibility (with frail population groups) coincide with areas of high disease occurrence then there is likely to be less interest in these areas (in terms of adverse disease presence), than areas where there is high disease occurrence and low number of susceptibles. Local occurrence of disease (counting of cases within areas) within short time spans (e. g. individual months or years) is termed *incidence.* Longer term accumulation of disease cases is often termed *prevalence.* Here the term incidence is used throughout. In general prevalence can be analyzed as for incidence.

To simplify discussion, initially, assume there is a small administrative area (such as a census tract, postcode, zip code, county etc.) within which it is possible to observe the disease incidence. Often there is a desire to compare the observed count of disease with what would have arisen from the underlying population. This will tell us if there is any excess disease risk in the local area. Let's assume there are $i = 1, \ldots, p$ tracts or small areas in a study area. Often a ratio of the observed count y_i, in the ith tract to the expected count e_{i} derived from the background population is used to examine excess risk: the *relative risk* of a disease within the i th area can be estimated by $\frac{y_i}{e_i}$. This ratio represents the relative risk compared to that the local population suggests should be seen in the area. Usually the count y_{i} will be available from government PH data sources and the expected count (or rate) is usually computed from known rates for the disease in population subgroups (broken by age and gender). This is known as *standardization*. The calculation of expected rates can be very important and different methods of calculation could lead to different conclusions about disease risk. Note that this relative risk definition implies a multiplicative model for risk. This is a common assumption in epidemiology.

Standardized Mortality/Morbidity/Incidence Ratio(SMR or SIR)

The above relative risk ratio is commonly computed for certain types of data. The most common is where incident cases are involved and is called the *standardized incidence ratio* (SIR). Sometimes live cases are described by the term morbidity and so SMR is sometimes used. This can be confusing as when deaths from a disease are recorded (mortality) the same acronym is applied (SMR: standardized mortality ratio). Of course, different expected rate calculations (denominators) would usually be used depending on whether incidence or mortality were to be considered.

Standardization Expected rates in the small areas or tracts $\{e_i\}$ are calculated (estimated) from the local population structure. Usually an external standard population rate will be known and applied to the local population. For example, suppose that the national US rate for prostate cancer (PrCa) is to be used to standardize the rates in South Carolina counties. The rate for different population groups must be known. Hence the rate for each age × gender group must be known nationally and also the population in this groups must be known locally. Define the US rate for PrCa in the kth age group and jth gender group as e_{kj}. Define the population in these groups in the ith area as p_{kji}. Hence the expected rate in the i th area will be simply:

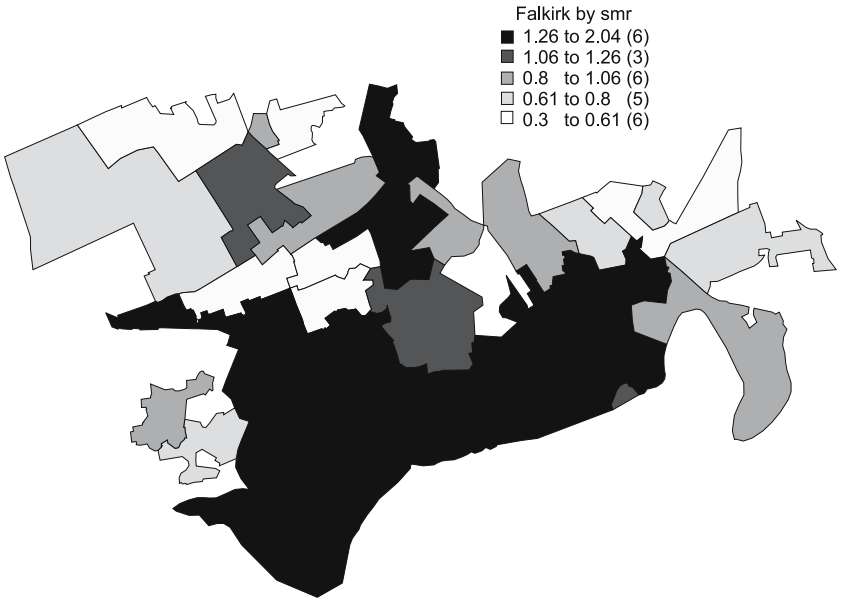
$$e_i = \sum_j \sum_k e_{jk} p_{kji} \,. \tag{1}$$

That is: the numbers in each tract in different age × sex groups are multiplied with known rates for the disease for equivalent groups in a *standard* population. The standard population may be the *national* population (as above) or even the *study region* population. The study region population may be the most relevant if you want to study relative spatial differences across a study region. Note also that other standardizations could be used where covariates are used to standardize the rates.

The standardized ratio of either incidence, mortality or morbidity is the relative risk ratio computed with standardized expected rates, as specified above:

$$SMR_i = \frac{y_i}{e_i} \,. \tag{2}$$

Figure 1 displays a standardized mortality ratio map for 26 census tracts in Falkirk central Scotland. The SMR map is often used by PH professionals to examine the distribution of disease risk. Areas of the map with SMRs greater than 2 or 3 (say) may be of concern. More formally, tests can be carried out to assess whether risk excesses are significant statistically. Visual assessment is not adequate for this purpose. Note also that the SMR is one estimate of relative risk, and there are many other ways to estimate risk.



Public Health and Spatial Modeling, Figure 1 Central Scotland: 26 census enumeration districts (EDs) in the center of the city of Falkirk. Respiratory cancer deaths (SMRs) for the period 1976–1983. Scottish national rate used for standardization

Control Diseases and Expected Rates

Expected rates are commonly used to allow for population effects when count data is observed. Count data is often available readily from government sources. However, for some purposes there is a need to examine the spatial distribution of cases at finer spatial resolutions. Commonly the residential address of cases is the finest level of resolution that can be found. Usually this is only relevant if a small geographic study region is examined. In this case the data form a spatial point process. As for count data, there is a need take population variation into account when examining risk at this spatial resolution level.

Expected rates are usually only available at aggregated geographic scales (census tracts or such like areas) and cant be used effectively at fine resolutions to control for population variations. An alternative is to use the incidence of a *control disease* within the study region. A control disease is matched closely to the risk structure of the disease of interest, but must not display the incidence effect under investigation. For example, live births could be used as a control for childhood leukemia in clustering studies. In that case the address locations of all births would be used as a population surrogate. This leads to two point processes: the leukemia case distribution and the live birth distribution. Of course, live birth is not a disease but in this case is a population indicator. Another example would be the use of residential addresses of coronary heart disease (CHD) as a control disease for respiratory cancer in studies of air pollution. CHD would be thought to affect the same age structure as respiratory cancer, but may be unaffected by air pollution hazards. Of course this may not be a good control as it could be affected by lifestyle variables such as smoking. Lower body cancers (testes, penis, ovaries etc.) has been proposed as a control disease for respiratory tract diseases. These are less affected by inhilatory insults. However care must be taken as some air pollutants can target lower body organs (e. g. nickel). In addition the time lag of disease expression (which is long in the case of cancers) should also be considered.

Note that this control disease is a *geographical* control and is not matched to specific cases. The common feature of each control disease is that it shouldn't be related to the effect of interest. There is some debate about use of these controls as opposed to expected rates from external sources.

The Ecological and Atomistic Fallacy

Many mapping studies attempt to relate incidence of disease in regions with some other measurable *explanatory* variable relating to the etiology of the disease e. g. it may be necessary to examine the relation between the number of smokers in regions and the incidence of respiratory cancer in the same regions. This might be achieved by applying regression analysis to the disease incidence and explanatory variable. The relation between these variables will be statistical and may suffer from the fact that regional totals or averages are used in the assessment of the relation. Hence an average relationship can only be measured. There is no direct link between whether an individual smokes and whether they develop lung cancer.

The *ecological fallacy* arises when such regional average characteristics are ascribed to *individuals* within the region

concerned. Any region-based analysis will suffer from this problem. It is known in fact that in some extreme cases the relation between the covariate and the outcome is reversed when individual analysis is carried out. Hence, ecological analyses are some times viewed with caution. Of course, at the aggregate level the relation remains valid. The *atomistic fallacy* occurs when analysis is based on individuals, and the variability of individuals' response to disease is not accounted for in inference at the regional level. These, and other aggregation issues, are further discussed in [15,31,36].

Confounders and Deprivation Indices

All disease maps contain the influence of variables affecting, or pertaining to, the local population which are not accounted for in standardized rates or control diseases. There are two ways to try to allow for these effects:

1. *include as many known explanatory variables in the expected rate or regression model to allow for these effects.* (These variables are called known confounders.)
2. *include the effect of unmeasured confounders via the use of random effects.*

In the first case the solution is to include in the study as many known variables that affect the outcome so that extra variation is explained. Of course it may not be feasible to include all know confounders simply due to (realistic) study limitations. To make allowance for unmeasured confounders (whether known or unknown) it is possible to admit random effects into any regression models. These are additional unobserved variates that will soak up extra variation of various kinds.

Often adverse disease incidence is known to be related to a range of poverty-related explanatory variables e. g. unemployment, housing type,welfare status, car ownership. That is, measurable adverse risk is expected in areas where these variates indicate low income and poverty. These variables are often available from national census. There has been some effort to combine such variables in composite measures known as *deprivation indices* [6]. In North America these are often termed urbanicity indices. Deprivation indices are now routinely available from government census data organizations and can be incorporated directly into a disease map as a covariate or as an offset term.

Some Spatial Statistical Issues

A fundamental feature of geo-referenced data available for analysis in Public Health applications is that it is usually discrete (either in the form of a point process or counting process), and the cases of concern arise from within a local human population which varies in spatial density and in susceptibility to the disease of interest. Hence any model or test procedure must make allowance for this background (nuisance) population effect. The background population effect can be allowed for in a variety of ways. For count data it is commonplace to obtain *expected* rates for the disease of interest based on the age-sex structure of the local population (see e. g. [9], chap. 3), and some crude estimates of local relative risk are often computed from the ratio of observed to expected counts (e. g. standardized mortality/incidence ratios: *SMR*s). For case event data, expected rates are not available at the resolution of the case locations and the use of the spatial distribution of a control disease has been advocated. In that case the spatial variation in the case disease is compared to the spatial variation in the control disease. A major issue in this approach is the correct choice of control disease. It is important to choose a control which is matched to the age-sex structure of the case disease but is unaffected by the feature of interest. For example, in the analysis of cases around a putative health hazard, a control disease should not be affected by the health hazard. Counts of control disease cases could also be used instead of expected rates when analyzing count data.

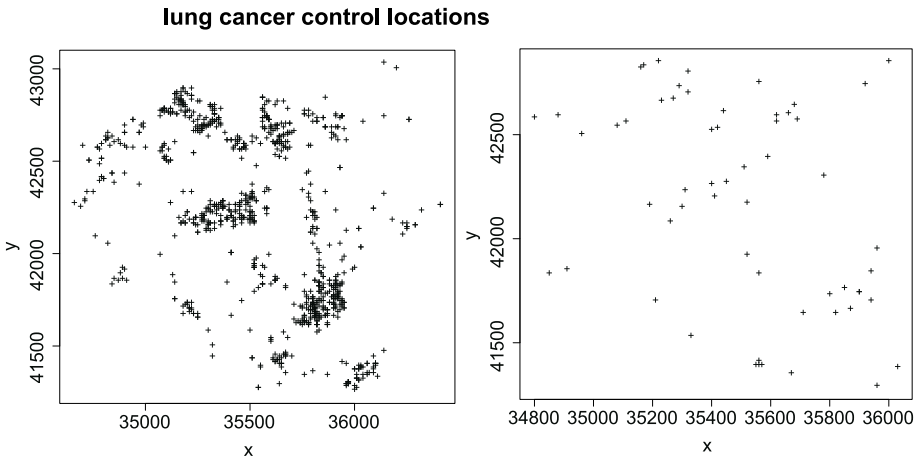
Case Event Data

Figure 2 displays control event (a) and case data (b) maps for a region of the UK for a fixed time period. In this example, larynx cancer case incidence is the distribution of interest while lung cancer distribution is the control disease for the same period.

Case event locations often represent residential addresses of cases and the cases arise from a heterogeneous population that varies both in spatial density and in susceptibility to disease. A heterogeneous Poisson process model is often assumed as a starting point for further analysis. Define the *first-order intensity* function of the case event process as $\lambda(\mathbf{s})$, representing the mean number of events per unit area in the neighborhood of location $\mathbf{s}$. This intensity may be parametrized as:

$$\lambda(\mathbf{s}) = \rho.\lambda_0(\mathbf{s}).\lambda_1(\mathbf{s};\theta) \tag{3}$$

where ρ is the overall rate of the process, $\lambda_0(\mathbf{s})$ is the 'background' intensity of the population at risk at $\mathbf{s}$, and $\lambda_1(\mathbf{s};\theta)$ is a parametrized function of risk. The focus of interest for making inference regarding parameters describing excess risk lies in $\lambda_1(\mathbf{s};\theta)$,, treating $\lambda_0(\mathbf{s})$ as a nuisance function. The function $\lambda_1(\mathbf{s};\theta)$ represents the relative risk measured locally around location $\mathbf{s}$, and $\log\lambda_1(\mathbf{s};\theta)$ is often modeled.



Public Health and Spatial Modeling, Figure 2 Lancashire UK: **a** lung cancer control disease address locations, **b** larynx cancer address locations. Both maps are for the period 1974–1983 and are for incident cases

It is possible that population or environmental heterogeneity may be unobserved in the data set. This could be because either the population background hazard is not directly available or the disease displays a tendency to cluster (perhaps due to unmeasured covariates). The heterogeneity could be spatially correlated, or it could lack correlation in which case it could be regarded as a type of *overdispersion*. One can include such unobserved heterogeneity within the framework of conventional models as a random effect.

This approach can lead to maximum a posteriori estimators similar to those found for universal kriging in geostatistics [23].This approach can also be implemented in a fully Bayesian setting (see e. g. [25] amongst others).

Count Data

Figure 3 displays a typical count data example: congenital death counts for South Carolina counties for the year 1990. A considerable literature has developed concerning the analysis of count data in spatial epidemiology (e. g. see reviews in [3,9,25,26]).

The usual model adopted for the analysis of region counts is to assume that $\{y_i, i = 1, \ldots, p\}$ are independent Poisson random variables with parameters $\{\lambda_i, i = 1, \ldots, p\}$. Here

$$\lambda_i = \int_{W_i} \lambda(\mathbf{s})\, \mathrm{d}\mathbf{u}\,, \quad i = 1, \ldots, p\,,$$

where $\lambda(\mathbf{s})$ is the first order intensity of the underlying cases and W_i is the i-th subregion. Often the $\lambda_i s$ are assumed to be constant within areas. Usually the expected count is modeled as

$$E(y_i) = \lambda_i = e_i \theta_i\,, \quad i = 1, \ldots, p\,.$$

This model may be extended to include unobserved heterogeneity between regions by introducing a prior distribution for the log relative risks ($\log \theta_i, i = 1, \ldots, p$). Incorporation of such heterogeneity has become a common approach and the Besag, York and Mollié (BYM) convolution model is now a standard model [29]. A full Bayesian analysis using this model is available on WinBUGS (available free from www.mrc-bsu.cam.ac.uk/bugs).

Key Applications

Disease Mapping

In this area, focus is on the processing of the disease map to take out random noise. Often applications in health services research require the production of an 'accurate' map of relative risks. Models for relative risk range from simple SMRs to posterior expected estimates from Bayesian models. In the count data situation, define the model for the observed counts as

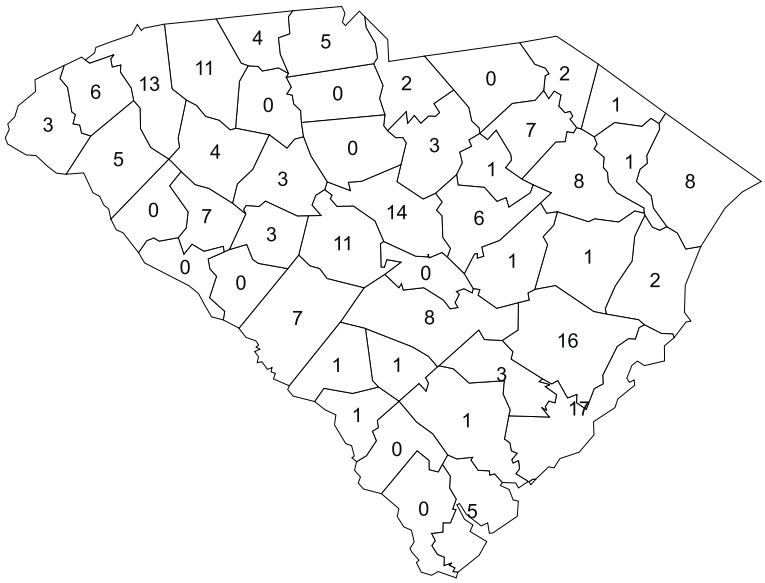
$$y_i \sim \mathrm{Poisson}(e_i \theta_i)$$
$$\log \theta_i = \mathbf{x}_i^\mathrm{T} \beta + \text{random effect terms}\,,$$

where $\mathbf{x}_i^\mathrm{T}$ is the i th row of a covariate design matrix and **\beta** is a regression parameter vector.

The simplest model assumes no linkages to covariates or random terms and the ML estimator of θ_i is the SMR: i. e. $\widehat{\theta_i} = y_i / e_i$. More often, and more generally, $\log \theta_i$ is assumed to be equal to a linear predictor involving covariates and regression parameters ($\mathbf{x}_i^\mathrm{T} \beta$). The final extension includes random effect terms to allow for overdispersion (uncorrelated heterogeneity UH: v_i) and spatially-correlated heterogeneity (CH: u_i). The model would then take the form:

$$\log \theta_i = \mathbf{x}_i^\mathrm{T} \beta + v_i + u_i\,.$$

In applications without covariates, when simple smoothing of rates is required, a simpler random effect model would



Public Health and Spatial Modeling, Figure 3
South Carolina: counts of congenital deaths by county in 1990

be used:

$$\log \theta_i = v_i + u_i \,.$$

This model assigns noise to two components: UH and CH. Both components are usually fitted to capture all the noise components thought to be present. This is often termed the BYM convolution model. To be able to estimate these components, prior distributions are assumed for each component. Usually these consist of an uncorrelated zero mean normal distribution for the overdispersion:

$$v_i \sim N(0, \tau_v) \,,$$

where τ_v is a variance parameter, and a spatial correlation prior distribution for the CH component. This could be chosen in a variety of ways. Commonly a Markov random field (MRF) is assumed. The intrinsic singular Gaussian distribution ([21,5,30]) is used where the conditional mean of the region effect is based only on a neighborhood of the region:

$$[u_i|.....] \propto N(\overline{u}_{\delta_i}, \tau_u/n_{\delta_i})$$

where δ_i is a neighborhood of the i th area, and n_{δ_i} is the number of regions in the i th neighborhood, and τ_u is a variance parameter which controls the degree of smoothing. This can be sampled within a posterior distribution sampling algorithm relatively simply. One alternative to this specification is to assume a fully parametrized covariance and a Multivariate normal distribution for CH:

$$\mathbf{u} \sim \mathbf{N}_p(\mathbf{0}, \Sigma)$$

where the elements of Σ are $\sigma_{ij} = cov(u_i, u_j)$. These covariance elements can be parametrized with a distance-based form such as $\sigma_{ij} = c_0 + \tau \exp(-\alpha d_{ij}^{\psi})$. Here a sill and nugget effect are specified and at zero distance the instantaneous variance is $c_0 + \tau$. This is more heavily parametrized than the MRF model above and the model also requires the inversion of a $p \times p$ covariance matrix. This of course allows for more detailed covariance modeling.

In a full Bayesian analysis, all parameters (β, $\mathbf{u}$, $\mathbf{v}$, $\tau_* \ldots$) would be assigned prior distributions and posterior sampling of these parameters, usually via MCMC algorithms, would be required.

For case event data, point process models must be considered initially. A heterogeneous Poisson Process model could be considered for p case events $\{\mathbf{s}_i\}$ $i = 1, \ldots, p$. It is possible to extend such a model to deal with random effects also. However when a control disease is also available, then it is possible to consider a simpler conditional logistic analysis. Define the joint realization of p cases and q controls as $i = 1, \ldots, p$ for the cases and $i = p+1, \ldots, p+q$ for the controls. Assume that the first order intensity of the cases is $\lambda(s, \theta) = \rho\lambda_0(s, \theta)\lambda_1(s, \theta)$ and of the controls $\lambda_0(s, \theta)$. Define the binary indicator variable y_{i} as follows:

$$y_i = \begin{cases} 1 & \text{if } \mathbf{s}_i \text{ is a case} \\ 0 & \text{otherwise} \end{cases}$$

then the conditional probability of a case at s_i is just

$$\frac{\rho\lambda_1(\mathbf{s}_i, \theta)}{1 + \rho\lambda_1(\mathbf{s}_i, \theta)} \,.$$

Hence, the likelihood of the realization is a logistic likelihood ([8]) specified by

$$L(\theta|\{\mathbf{s}_i\}) = \prod_{l=1}^{p+q} \frac{[\rho\lambda_1(\mathbf{s}_i,\theta)]^{y_i}}{1+\rho\lambda_1(\mathbf{s}_i,\theta)}\,. \tag{4}$$

A suitable specification for the relative risk $\lambda_1(\mathbf{s}_i,\theta)$ could be $\log\lambda_1(\mathbf{s}_i,\theta) = \mathbf{x}_i^{\mathrm{T}}\beta + v_i + u_i$ where any covariates would have to be available at all case and control locations. Note that a model without covariates only requires random effect estimates at locations. Specifying suitable prior distributions for such a model is not difficult and, for example, first order neighborhoods of points can be obtained from tessellation information ([2]), and so MRF prior distributions can be specified. Alternative semi-parametric models have been suggested by [17].

Disease Clustering

In this area, the focus is not on reduction of noise, per se, but the assessment of the clustering tendency of the map and in particular the assessment of which areas of a map display clustering. Here, clustering could be around a known putative source of hazard (*focused* clustering) or have no known locations of clustering (*non-focused* clustering). A variety of testing methods are available for cluster detection, see for example [19].
However it is also possible to consider modeling clusters. In general the model formulation may not differ greatly from that of relative risk estimation, depending largely on the definition of clusters and clustering.

Focused Clustering Focused clustering is the simplest case and usually assumes that some form of distance decrease in risk happens around a fixed point or points.
For example, the count data model can be defined as

$$y_i \sim \text{Poisson}(e_i\theta_i)$$
$$\log\theta_i = \log(1+\exp\{-\alpha d_i\}) + \mathbf{x}_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}\gamma\,,$$

where d_{i} is a distance measured to the small area from the focus point (such as a chimney, mobile phone mast, or waste dump site). Here the extra covariates appear in $\mathbf{x}_i^{\mathrm{T}}$ while the z_i^{T} is the i th row of a matrix of random effects and γ is a unit vector. In this case focus is on inference concerning α as this defined the distance relation. Within $\mathbf{x}_i^{\mathrm{T}}$ there could also be directional terms such as $\cos(\phi)$and $sin(\phi)$,where ϕ is the angle between the area (centroid) and the focus point. This can be used to detect any directional concentration of risk (which could be important particularly if an air pollution risk is possible).
For case events, the case event locations are often assumed to follow a heterogeneous Poisson process with first order intensity $\lambda(\mathbf{s})$. Denote this as $\{\mathbf{s}_i\} \sim PP(\lambda(\mathbf{s}))$.If a control disease is available and the conditional logistic likelihood (4) is assumed then the intensity can be parametrized as:

$$\log\lambda_1(\mathbf{s}_i,\theta) = \log(1+\exp\{-\alpha d_i\}) + \mathbf{x}_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}\gamma\,,$$

where d_{i} is the distance from any case or control event to the focus point. Directional effects can be included here also as for count data. When fixed effects are included only with no covariates, then a frequentist approach would allow the estimation of α via maximum likelihood. Equally, if a Bayesian approach is assumed then all parameters would have prior distributions and the resulting posterior distribution would usually be sampled. Some general references for this area are [9 chap. 9,25 chap. 7].

Non-Focused Clustering When locations of clusters are unknown then the statistical task becomes more difficult. Not only are the locations of putative clusters unknown but their number and size are also not predefined. This area can be further divided into *general* clustering, where the overall tendency of an area to cluster is assessed, and *specific* clustering where the locations of clusters are to be assessed. Many testing procedures have been derived to assess general clustering tendency (see e. g. for case events: [1,7]; and for counts: [20,34]). Fewer procedures are available for specific clustering. Scan statistics (SatScan) have been proposed ([19]).
Modeling of clusters can be approached in a variety of ways. First, if clustering of excess risk is simply and liberally regarded as *significant excess risk found anywhere on a map* then pointwise determination of excess can be pursued. This is known as hot-spot clustering. For example, for count data, it could be assumed:

$$y_i \sim \text{Poisson}(e_i\theta_i)$$

as before, and examine either *i)* estimates of θ_i for unusual features (usually significantly elevated values), or, *ii)* the residuals from a fitted model:

$$\widehat{r_i} = y_i - e_i\widehat{\theta_i}$$

to find out if, after model fitting, whether there are areas of excess unexplained by the model.
The first approach assumes a model for risk and under that model some form of cluster identification may take place. Alternatively a model which simply cleans noise out may be considered i. e. a model for $\log\theta_i$ is assumed such as $\log\theta_i = \mathbf{x}_i^{\mathrm{T}}\beta + v_i$. This model allows for covariate adjustment and some extra variation but does not model CH (smoothing) as this may reduce its ability to detect

aberrations in risk at the single region level. Following the model fit an assessment of the significance of $\widehat{\theta_i}$ could be made.

If on the other hand a specific structure for clusters is assumed then a formal clustering model may be assumed. There is a gray area between relative risk estimation (which focuses on the estimation of θ_i) and i) above where estimates of θ_i are examined for significant excess. If some form of cluster identification is included in the model then that can be checked for location and size of clusters. This can be useful when data are sparse and other global CH models cant describe the cluster form. One proposal is for risk to be related to a set of hidden (unobserved) cluster locations:

$$\log \theta_i = \mathbf{x}_i^{\mathrm{T}} \beta + v_i + \log\{1 + \sum_{k=1}^{K} h(\mathbf{x}_i; \xi_k)\}$$

where there are K unknown clusters with locations $\{\xi_k\}$, $\mathbf{x}_i$ is the centroid of the i th small area and $h(\mathbf{x}_i; \xi_k)$ is a cluster distribution function that describes the relation of any point to a cluster location. Usually $h(\mathbf{x}_i; \xi_k)$ is designed to have a decline in risk with distance from ξ_k, but a range of forms are available. Unfortunately, given that K is unknown, a number of assumptions must be imposed on the analysis to allow for estimation of parameters. Often reversible jump McMC is employed here (see e. g. [13,27,28]). For case event data, models for $\log \lambda_1(\mathbf{s}_i, \theta)$ can be set up with similar considerations (see e. g. [24]).

The second approach, that of examining residuals such as $\widehat{r_i} = y_i - e_i\widehat{\theta_i}$, may be useful if a noise reduction model is used in the estimation of $\widehat{\theta_i}$. However the residual will always include some form of noise unrelated to clustering. Even a perfect model will always have Poisson noise around the true risk: $e_i\theta_i$. Hence it would be important to specify the risk model carefully to allow for only clustering effects to appear in the residual as far as possible. 'Unusual' residuals can be examined via Monte Carlo procedures such as parametric Bootstrap or, under a Bayesian paradigm, a Bayesian Bootstrap using the predictive distribution.

Finally alternative approaches that assume that clusters are defined within areas or neighborhoods (as opposed to single regions) can be considered and diagnostics for these have been proposed [16].

Ecological Analysis

In this area, the relation between disease incidence and explanatory variables is the focus, and this is usually carried out at an aggregate level, such as with counts in small areas.

Many issues of bias and misclassification error can arise with ecological data and the interested reader is referred to [36] and [14] for further insights.

Two important areas of concern are related to scale aggregation issues: MAUP and MIDP. The Modifiable areal Unit Problem (MAUP) concerns the scalability of models and whether, at different spatial scales, a model is valid. In general this is unlikely to be the case as far as covariance structure is concerned as this would lead to fractal covariances which are not found commonly. However, the labeling of scales of relevance of models is important and the extent to which a model can be scaled is relevant in many applications. A related but different issue is how to use different scales of data within one analysis i. e. should individual level data be used in preference to aggregated data or can they be combined. This is a focus of current research.

The misaligned data problem (MIDP) is related to the last issue, but specifically addresses the issue of combining data from different spatial scales to provide analysis at one level. For example health outcomes (disease incidence etc.) may be observed within census tracts and there are available pollution measurements at monitoring sites around the study area. To make inferences about the health outcomes it is best to use the pollution data relevant to the census tracts. One simple solution would be to block Krige the pollution data to provide block estimates for each of the tracts (see e. g. [3, chap. 6,32]). This would ignore the error in the interpolation of the pollution data of course and a better approach is to consider a model where the true exposure is modeled within the health model but the pollution model is jointly estimated.

The model often assumed for count data is of the form

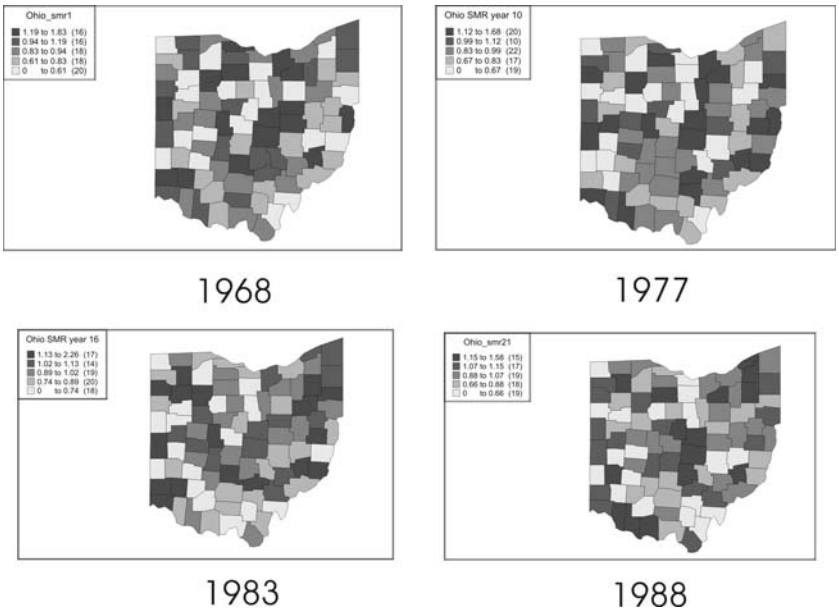
$$\begin{aligned} y_i &\sim \text{Poisson}(e_i\theta_i) \\ \log \theta_i &= \mathbf{x}_i^{\mathrm{T}} \beta + z_i^{\mathrm{T}} \gamma \,. \end{aligned}$$

Assuming that it is possible to observe count data (y_i) and also observe measurements $\{x(\mathbf{s}_j)\}$ made at q sites. Assume the measures have mean $E(x(\mathbf{s}_j)) = \mu(\mathbf{s}_j)$, and they are multivariate normal with covariance $\text{cov}(x(\mathbf{s}_i), x(\mathbf{s}_j)) = \sigma_{xij}$. Also Σ is the covariance matrix with ijth element σ_{xij}. For a block, the mean is defined as $\mu_{B_i} = |B_i|^{-1} \int_i \mu(\mathbf{s}_j)\, \mathbf{du}$, where B_i denotes the ith area, and an estimate is $\widehat{\mu}_{B_i} = |B_i|^{-1} \int_i \widehat{\mu}(\mathbf{s}_j)\, \mathbf{du}$. It could be assumed in this case:

$$\begin{aligned} \{y_i\} &\overset{ind}{\sim} \text{Poisson}(e_i\theta_i) \\ \log \theta_i &= \beta\mu_{B_i} + z_i^{\mathrm{T}} \gamma \,. \end{aligned}$$

and jointly with

$$\{x(\mathbf{s}_j)\} \sim N_q(\mu(\mathbf{s}_j), \Sigma) \,,$$



Public Health and Spatial Modeling, Figure 4 Ohio county map: respiratory cancer SMRs for 4 selected years: 1968,1977,1983, and 1988

μ_{B_i} can be estimated and the associated error can be accounted for. Similar considerations can apply to case event data.

Space-Time Modeling and Disease Map Surveillance

Space-Time Models The extension of mapping models to space-time is straightforward in the case of counts within areas within time periods. Figure 4 displays sequences of maps of respiratory cancer for 4 year periods in the counties of the US State of Ohio. Space-time variation in risk is apparent from the variation from year to year for given counties.

For example, yearly counts of disease within small areas can be handled relatively straightforwardly. In this area, the focus is the construction of methods which, usually, examine the spatio-temporal variation of disease. A typical count data model (for counts y_{ij} in the ith region and jth time period) might be

$$y_{ij} \sim \text{Poisson}(e_{ij}\theta_{ij})$$
$$\log \theta_{ij} = \alpha + (\text{covariates}) + u_i^{\mathrm{T}}\gamma + w_j^{\mathrm{T}}\xi + z_{ij}$$

where $u_i^{\mathrm{T}}\gamma$ is a sum of spatial random components γ is a unit vector), and $w_j^{\mathrm{T}}\xi$ is a sum of temporal effects ξ is also a unit vector) and z_{ij} is a space-time interaction effect. This formulation can lead to a rich variety of models depending on the definition of the structure of the components. [18] discusses various possibilities in the Bayesian context.

Map Surveillance Surveillance usually requires there to be a prospective view taken of the data whereby new events are recorded and detection of changes in the vent pattern is important. This may be done in real- or near-real-time. This area has become important due to bioterrorism threats and the possibility of large scale PH disaster prediction. Often a space-time model must be general enough to cope with normal variation in risk but also capable of detecting aberrations as they arise. One useful approach is to consider the predictive distribution of data given previous events and compare this with the new events. This leads to so called *surveillance residuals* [35].

Certain optimal methods are available for the detection of changes in a disease incidence and clustering in space-time (see [11,12,22] provides reviews). General methods for detecting temporal disease changes are given in [10].

Future Directions

There are many open problems in this area. While much attention has been paid to putative hazard assessment (focused clustering) and also methods for relative risk estimation, there is still considerable need for development of methodology for cluster detection and also multi-focus surveillance in real-time. Future directions will see the development of multivariate models and also the fuller examination of space-time.

Cross References

- ► Autocorrelation, Spatial
- ► Biomedical Data Mining, Spatial
- ► Hierarchical Spatial Models
- ► Homeland Security and Spatial Data Mining
- ► Spatial and Geographically Weighted Regression

► Spatial Regression Models
► Statistical Descriptions of Spatial Patterns

Recommended Reading

1. Anderson, N.H., Titterington, D.M.: Some methods for investigating spatial clustering, with epidemiological applications. J. R. Stat. Soc. **160**, 87–105 (1997)
2. Baddeley, A., Turner, R.: Spatstat: An r package for anayzing spatial point patterns. J. Stat. Softw. **12**, 1–42 (2003)
3. Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC Press, London (2004)
4. Bernardinelli, L., Clayton, D.G., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M.: Bayesian analysis of space-time variation in disease risk. Stat. Med. **14**, 2433–2443 (1995)
5. Besag, J., York, J., Mollié, A.: Bayesian image restoration with two applications in spatial statistics. Ann. Inst. Stat. Math. **43**, 1–59 (1991)
6. Carstairs, V.: Small area analysis and health service research. Community Med. **3**, 131–139 (1981)
7. Diggle, P.J., Chetwynd, A., Haggvist, R., Morris, S.: Second-order analysis of space-time clustering. Stat. Methods Med. Res. **4**, 124–136 (1995)
8. Diggle, P.J.: Point process modellng in environmental epidemiology. In: Barnett, V., Turkman, K.F. (eds.) Statistics in the Environment SPRUCE I. Wiley, New York (1993)
9. Elliott, P., Wakefield, J.C., Best, N.G., Briggs, D.J. (eds.): Spatial Epidemiology: Methods and Applications. Oxford University Press, London (2000)
10. Farrington, P., Andrews, N.: Outbreak detection: Application to infectious disease surveillance. In: Brookmeyer, R., Stroup, D. (eds.): Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance. chapter 8. Oxford University Press, London (2004)
11. Frisen, M.: Statistical surveillance: optimality and methods. Int. Stat. Rev.. **71**, 1403–1434 (2003)
12. Frisén, M., Sonesson, C.: Optimal surveillance. In: Lawson, A.B., Kleinman, K. (eds.) Spatial and Syndromic Surveillance for Public Health, chapter 3. Wiley, New York (2005)
13. Gangnon, R.: Impact of prior choice on local Bayes factors for cluster detection. Stat. Med. **25**, 883–895 (2006)
14. Gustafson, P.: Measurement Error and Misclassification in Statistics and Epidemiology. Chapman & Hall, London (2004)
15. Haneuse, S., Wakefield, J.: Ecological inference incorporating spatial dependence. In: King, G., Rosen, O., Tanner, M. (eds.) Ecological Inference: New Methodological Strategies, chapter 12, pp 266–301. Cambridge University Press, London (2004)
16. Hossain, M., Lawson, A.B.: Cluster detection diagnostics for small area health data: with reference to evaluation of local likelihood models. Stat. Med. **25**, 771–786 (2006)
17. Kelsall, J., Diggle, P.: Spatial variation in risk of disease: a nonparametric binary regression approach. Appl. Stat. **47**, 559–573 (1998)
18. Knorr-Held, L.: Bayesian modelling of inseparable space-time variation in disease risk. Stat. Med. **19**, 2555–2567 (2000)
19. Kulldorff, M., Nagarwalla, N.: Spatial disease clusters:detection and inference. Stat. Med. **14**, 799–810 (1995)
20. Kulldorff, M., Tango, T., Park, P.J.: Power comparisons for disease clustering tests. Comput. Stat. Data Anal. **42**, 665–684 (2003)
21. Kunsch, H.R.: Intrinsic autoregressions and related models on the two-dimensional lattice. Biometrika. **74**, 517–524 (1987)
22. Lai, T.L.: Sequential changepoint detection in quality control and dynamical systems. J. R. Stat. Soc. B **57**, 613–658 (1995)
23. Lawson, A.B.: On using spatial Gaussian priors to model heterogeneity in environmental epidemiology. The Statistician. **43**, 69–76 (1994). Proceedings of the Practical Bayesian Statistics Conference
24. Lawson, A.B.: Cluster modelling of disease incidence via rjmcmc methods: a comparative evaluation. Stat. Med. **19**, 2361–2376 (2000)
25. Lawson, A.B.: Statistical Methods in Spatial Epidemiology, 2nd edn. Wiley, New York (2006)
26. Lawson, A.B., Böhning, D., Lessafre, E., Biggeri, A., Viel, J.F., Bertollini, R. (eds.) Disease Mapping and Risk Assessment for Public Health. Wiley, Chichester (1999)
27. Lawson, A.B., Clark, A.: Markov chain Monte Carlo methods for clustering in case event and count data in spatial epidemiology. In: Halloran, M.E., Berry, D. (eds.) Statistics and Epidemiology: Environment and Clinical Trials, pp 193–218. Springer, New York (1999)
28. Lawson, A.B., Denison, D.: Spatial cluster modelling: an overview. In: Lawson, A.B., Denison, D. (eds.) Spatial Cluster Modelling, chapter 1, pp. 1–19. CRC press, New York (2002)
29. Mollié, A.: Bayesian and empirical bayes approaches to disease mapping. In: Lawson, A., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F., Bertollini, R. (eds.) Disease Mapping and Risk Assessment for Public Health, chapter 2, pp. 15–29. Wiley, Chichester (1999)
30. Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC, New York (2005)
31. Salway, R., Wakefield, J.: Sources of bias in ecological studies of non-rare events. Environ. Ecol. Stat. **12**, 321–347 (2005)
32. Schabenberger, O., Gotway, C.: Statistical Methods for Spatial Data Analysis. Chapman & Hall, London (2004)
33. Snow, J.: On the Mode of Communication of Cholera. Churchill Livingstone, London, 2nd edn. (1854)
34. Tango, T.: Comparison of general tests for spatial clustering. In: Lawson, A.B., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.F., Bertollini, R. (eds.) Disease Mapping and Risk Assessment for Public Health, chapter 8. Wiley, New York (1999)
35. Vidal-Rodeiro, C., Lawson, A.: Monitoring changes in spatio-temporal maps of disease. Biom. J. **48**, 1–18 (2006)
36. Wakefield, J.: A critique of statistical aspects of ecological studies in spatial epidemiology. Environ. Ecol. Stat. **11**, 31–54 (2004)

PVD

► Floating Car Data

Pyramid Technique

CHRISTIAN BÖHM
Institute for Computer Science Database and Information Systems, University of Munich, Munich, Germany

Synonyms

Pyramid tree

Definition

The *Pyramid Technique* [1] is an indexing technique for point data (feature vectors) of a multidimensional space, particularly designed for medium to high dimensionality starting from $d = 10$. Like Z-ordering [2] and other space-filling-curve techniques the pyramid technique gives a one-dimensional embedding of the high dimensional points. The embedded objects can be indexed by any one-dimensional index structure which supports range queries (interval queries) such as all B-tree [3] variants as well as all order preserving hashing methods. The pyramid technique can efficiently handle multidimensional interval queries and nearest neighbor queries using maximum metric.

Historical Background

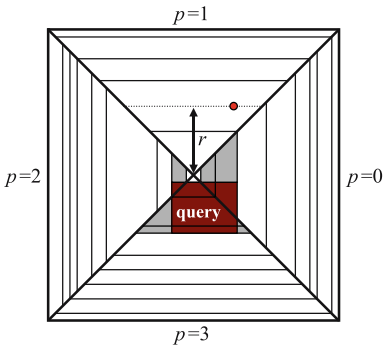
Index structures for vector spaces of medium to high dimensionality [4,5,6] have become very popular in the 1990s, because traditional index structures for vector data, such as the R-tree [7] and its variants tend to deteriorate as the dimensionality of the space increases, an effect commonly referred to as the *curse of dimensionality* [8]. One-dimensional embedding techniques [2], in general, yield the advantage that the complete storage management is handled by an index structure for a one-dimensional space, which is readily available by commercial database systems. Therefore, features such as transaction processing, concurrency and recovery are inherited from the one-dimensional index. Although vector data are transformed into one-dimensional spaces, the complete technique is subject to the *curse of dimensionality*. Therefore, it is important to develop the pyramid technique, which is particularly designed for higher-dimensional spaces and suffers its problems to a lesser extent. The pyramid technique is, in general, not limited to a particular metric, but the schema of space partitioning makes it particularly suited for queries using the maximum metric. The pyramid technique has inspired a number of other techniques, such as the onion technique [9] or concentric hyperspaces [10] and many others, which focus on different metrics including the Euclidean metric.

Scientific Fundamentals

In contrast to most of the well-known space-filling curves, the pyramid technique does not rely on a *recursive* schema of space partitioning. In contrast, the data space is partitioned into $2 \cdot d$ hyper-pyramids which share the origin of the data space (which can be chosen as the center point of the data set) as top point and have each an individual $(d–1)$-dimensional basis area (cf. Fig. 1). The pyramids are systematically numbered which forms the first part of the embedding key (a natural number p). The second part is the distance (with respect to the maximum metric) from the origin (a positive real number r). The embedding key can be formed as an ordered pair $k = (p, r)$, or, equivalently, if the maximum of all r-values (r_{max}) is known, we can form one single embedding key $k' = r_{max} \cdot p + r$.

In both cases, a d-dimensional range query can be translated in a set of search intervals on the search keys. The number of intervals is at most $2 \cdot d$, because the query object can at most have one intersection with each of the pyramids (cf. Fig. 1. Since nearest neighbor queries can be transformed into range queries (which requires a set of at most two one-dimensional ranking processes per pyramid), it is also possible to evaluate nearest neighbor queries.

The *Extended Pyramid-Technique* was proposed to handle data with skewed data distribution. The idea is a translation of the the data set such that the center of the set is located at the reference point $(0.5, \ldots, 0.5)$ where all the pyramids of the original pyramid techniques share their top point while keeping the data in the unit hypercube $[0..1]^d$. Since the centroid (means) of the data points is not very stable in the presence of skewed data distributions, our technique is based on a median method which determines a point, which is the coordinate-wise median of all data points. We refer to this point as the median point $mp = (mp_1, \ldots, mp_d)$. Then we determine for each coordinate an expo-



Pyramid Technique, Figure 1 Space partitioning of the pyramid technique

nent r_i such that the following condition holds:

$$r_i = -\frac{1}{\log_2(mp_i)}.$$

This exponent is used in a function $t(x) = (x_1^{r_1}, \ldots, x_d^{r_d})$ to transform the data in a new space from which it has the nice property to map the points $[0..1]^d$ into the same space $[0..1]^d$, while moving the median point *mp* to the reference point (0.5, ..., 0.5) of the pyramids. The index must be rebuilt whenever the change of the median point *mp* extends a threshold. It was shown that this happens rarely. In addition to the general advantages of the one-dimensional embedding, the experimental evaluation of the pyramid technique and the extended pyramid technique yielded a considerable speed-up factor of up to 14 with respect to the number of page accesses, of up to 103 with respect to CPU consumption and of up to 2500 with respect to the overall response time over the X-tree.

Key Applications

High dimensional indexing is important for similarity search systems in various application areas, such as multimedia, CAD, systems biology, medical image analysis, time sequence analysis and many others. Complex objects are typically transformed into vectors of a high-dimensional space (feature vectors), and the similarity search thereby translates into a range or nearest neighbor query on the feature vectors. High-dimensional feature vectors are also required for more advanced data analysis tasks such as cluster analysis or classification.

Future Directions

One-dimensional embedding would also be interesting for several new metrics, such as set metrics (multi-instance objects) or for uncertain and moving objects. Only few approaches exist [11] to support general metric spaces.

Recommended Reading

1. Berchtold, S., Böhm, C., Kriegel, H-P.: The pyramid-technique: towards breaking the curse of dimensionality, pp. 142–153, SIGMOD Conference, (1998)
2. Orenstein, J.A.: Redundancy in spatial databases. SIGMOD Conference pp. 295–305 (1989)
3. Bayer, R., McCreight, E.M.: Organization and maintenance of large ordered indexes, pp. 107–141, SIGFIDET Workshop (1970)
4. Lin, K-I., Jagadish, H.V., Faloutsos, C.: The TV-tree: An index structure for high-dimensional data. VLDB J. 3(4):517–542 (1994) [DBLP:journals/vldb/LinJF94]
5. White, D., Jain, R.,: Similarity indexing with the SS-tree. ICDE 516–523 (1996)
6. Berchtold, S., Keim, D.A., Kriegel, H-P.: The X-tree: An index structure for high-dimensional data. VLDB, 28–39 (1996)
7. Guttman, Aa.: R-Trees: A dynamic index structure for spatial searching. SIGMOD Conference 47–57 (1984)
8. Berchtold, S., Böhm, C., Keim, D.A., Kriegel, H-P.: A cost model for nearest beighbor search in high-dimensional data space. PODS 78–86 (1997)
9. Chang, Y.-C., Bergman, L.D., Castelli, V., Li, C.-S., Lo, M.-L., Smith, J.R.: The onion technique: indexing for linear optimization queries. SIGMOD Conference 391–402 (2000)
10. Ferhatosmanoglu, H., Agrawal, D., Abbadi, A.E.: Concentric hyperspaces and disk allocation for fast parallel range searching. ICDE 608–615 (1999)
11. Jagadish, H.V., Ooi, B.C., Tan, K.-L., Cui, Yu, Rui, Zhang: iDistance: An adaptive B+-tree based indexing method for nearest neighbor search. ACM Trans. Database Syst. 30(2):364–397 (2005)

Pyramid Tree

► Pyramid Technique