

Models for diffusion of innovations and Cellular Automata: an epidemiological application to pleural mesothelioma

Cinzia Mortarino

Abstract In the city of Casale Monferrato, the largest Italian factory that produced asbestos cement goods was active from 1907 to 1985. As a consequence, asbestos fibers scattered in the surrounding area and caused an enormous number of pleural mesotheliomas. Due to the very long latency of this disease, many subjects have not exhibited its symptoms yet. The aim of this paper is to model and predict the future evolution of the number of deaths due to this disease among residents in the area around that city. The model used here is built as an *aggregation* of the evolution rule of a Cellular Automaton that is assumed to pass through three steps: exposure, contamination, diagnosis. In that way, forecasts of the future evolution take into account the environmental conditions that changed in time because of different levels in plant activity. The model is fitted to diagnosis data starting from the earliest until the first months of 2009. Results show that deaths will not end until 2030, and that in the next two decades, almost 500 more subjects will be diagnosed with this disease.

Key words: Statistical modeling, Cellular Automata, Pleural Mesothelioma

1 Introduction

Asbestos is the major cause for pleural mesothelioma (PM). Its symptoms appear after a long latency (almost always greater than 30 years). The aim of this paper is to model and predict the future evolution of the number of deaths due to PM among residents in the area around Casale Monferrato, Italy. In this city, the largest Italian factory that produced asbestos cement goods was active from 1907 to 1985. Data used in this study start from the earliest diagnosis until the first months of 2009.

In this section, only a schematic description of the large body of literature con-

Cinzia Mortarino

Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121 Padova, Italy
e-mail: mortarino@stat.unipd.it

cerning this subject will be presented. Most of the papers refer to professional exposure to asbestos. Data aggregated at the regional or more often the national level are used. The most recurring method is an *age-period-cohort (APC) analysis* in order to estimate the risk of dying of PM at different ages and different periods. Predictions of the death toll are then formulated by applying the risks to plausible future population scenarios ([9], [8], [7]). With reference to these works, the main concerns stem from the reliability of predictions after 2015/2020, because they refer essentially to men born at the end of the '50s or after, for whom no data are yet available, and the possibility that part of the increase in death rates trend is due to the growing awareness of and diagnostic ability regarding PM. Moreover, most of these works do not take into consideration that asbestos was not used throughout the last century consistently with the same intensity (and was banned at the beginning of the '90s from most industrialized countries). At best, different trends that arise from qualitative considerations are attempted in some models in order to choose the one that fits better to data. This of course adds uncertainty to future predictions. An exception is represented by [5], where the Italian national asbestos consumption curve is used as a measure of exposure in the population. Since the stronger relationship is observed when consumption data are related to deaths that occurred 40 years later, this value is used to relate risk estimates obtained with APC analysis to the predicted evolution of the death trend. In [5], however, the relative risk of the last period (1995–1999) has to be assumed to hold for future periods, and that assumption cannot be tested.

Papers dealing with environmental or domestic exposure (i.e., people living or working in structures built with asbestos materials) are less common. In all of them, the aim is to estimate the incidence rate for people living close to an asbestos mine or to an asbestos goods factory and to decide whether the incidence rate is significantly greater than the average rate in the rest of the population. Previous papers do not provide forecasts for future diagnoses. Works of this type pertaining to Italy deal with the area of Casale M. The method is often a case-control study. In particular, results in [6] denote that the risk decreased rapidly as the distance from the factory increased, but at 10 km the risk was still 60% of its value at the source.

In this paper, the aim is to model the death toll among residents in the area in order to quantify the size of the *epidemic* that in the last decade has only begun to show its gravity. The innovations of this work stem from the data sources and the method used, which attempts to describe how exposure, measured through yearly asbestos quantity used in the production plant, is linked to the expression of PM.

2 Model

In the model used here, the definition of *contamination* is given (see below), and, differently from the definition of latency commonly used, here latency will mean the time (number of years) from *contamination* to diagnosis (instead of time from *beginning of exposure* to diagnosis). The reason for this choice is that we believe it is worthwhile to build a model that relies on Cellular Automata evolution, [1]. In particular, the definition of a transition rule *at the individual level* allows all the peculiarities of this disease's expression and its connection with time-dependent

environmental conditions to be taken into consideration. Since, however, data are available only *at the population level*, as yearly numbers of diagnosis, we proceed by “summing up” the individual transition rule, in order to obtain a population model that could be fitted to our data. The estimates of this model can be interpreted again at the individual level. The method used here is inspired by the procedure used in [2], although some differences are required to take into account a completely different application. In [2], the aim is to model the latent process of information diffusion. Information is there considered as the prerequisite for adoption (of market goods or services), and, at each time point, the size of the “acquainted network” represents the market potential that could be actually reached by adoptions (observable process). In this work, the latent process is given by the contamination (which is linked to environmental conditions changing in time). This is the prerequisite for developing PM¹ and, after a long period of time, symptoms appear (observable process).

Contamination is defined here as the beginning of the disease in latent form. This term should not be confused with exposure, which means only that the subject lived in the Local Health Authority (LHA) for at least one year. An exposed subject may never get contaminated. A contaminated subject will surely develop PM after a latency period whose length varies from 1 to K years (K is a parameter to be estimated). Let us denote with

$$c(i; t) = \begin{cases} 0 & \text{if subject } i \text{ is not contaminated at time } t \\ 1 & \text{if subject } i \text{ is contaminated at time } t \end{cases}$$

the contamination step of a subject. This is a non-observable datum.

Given i , the contamination process unfolds according to the following rule:

$$c(i; t+1) = c(i; t) + Be(p_c(t+1)) I_{\{c(i; t)=0\}}, \quad (1)$$

where $Be(\alpha)$ denotes a Bernoulli random variable of parameter α , I_E denotes an indicator function whose value is equal to 1 when the event E is true and 0 elsewhere, $p_c(t+1)$ is the contamination probability that has to be modelled as a function of the environmental conditions at time t . A simple solution might be

$$p_c(t+1) = \gamma A(t),$$

where γ is a parameter that has to be estimated and $A(t)$ is a the known time series of the yearly asbestos quantity used in the plant. Parameter γ represents the increase in contamination probability due to an extra asbestos ton worked in the plant in year t . In that expression, the contamination probability does not depend upon exposure duration (but on exposure intensity); nevertheless, an individual with longer exposure has a greater contamination probability because, for a greater number of time points t , the Bernoulli r.v. of Eq. (1) will be realized. This structure describes the consensus of most researchers agreeing there is a positive dose-response curve for PM—the stronger the exposure, the greater the risk ([3]). Moreover, this choice to model this process as a sequence of steps “exposure (E) – contamination (C) – diagnosis (D),” by denoting with latency the distance between C and D and by describing the C probability as an increasing function of exposure intensity is coherent with the

¹ Once inhaled, bigger asbestos fibers are eliminated while the smaller ones transfer to body tissue, where they become permanently lodged, and, in some subjects, begin the carcinogenesis process.

common belief that longer latencies arise for less intensive exposures ([5]): in the literature, latency measures the distance between E and D; a long latency (years from E to D) for light exposures may be the result of a low C probability, which on average leads to a longer time from E to C, i.e., a longer time from E to D.

Notice that factory's workers are not pulled separately from the rest of the population because the aim is to predict the death toll (of both workers and non-workers) and not to measure a subject's risk. The possibility of a different evolution rule for workers and non-workers might be a solution that will be studied in the future.

For the observable component (PM diagnosis), the following notation is used:

$$m(i;t) = \begin{cases} 0 & \text{if subject } i \text{ is not diagnosed with PM at time } t \\ 1 & \text{if subject } i \text{ is diagnosed with PM at time } t. \end{cases}$$

Given i , the PM expression's evolution rule is modelled as follows:

$$m(i;t+1) = m(i;t) + Be(p_{mi}(t+1))I_{\{m(i;t)=0\}} \quad (2)$$

where $p_{mi}(t+1)$ denotes the probability that subject i is diagnosed with PM at time $(t+1)$. Let Y denote the r. v. measuring the latency length. Y can assume values in $\{1, 2, \dots, K\}$. Let π_y , $y = 1, 2, \dots, K$ denote the probability of $\{Y = y\}$. For subject i contaminated from s years, $p_{mi}(t+1)$ equals the hazard rate of Y evaluated in s :

$$\begin{aligned} p_{mi}(t+1) &= P[m(i;t+1) = 1 \mid m(i;t) = 0, c(i;t-s+1) = 1, c(i;t-s) = 0] \\ &= P[Y = s \mid Y \geq s] = \pi_s / (\pi_s + \pi_{s+1} + \dots + \pi_K). \end{aligned}$$

A sensible assumption for $p_{mi}(t+1)$ might be

$$p_{mi}(t+1) = \sum_{k=0}^{K-1} c(i;t-k) \mu_{k+1}, \quad (3)$$

where μ_s , $s = 1, 2, \dots, K$, are appropriate parameters linking the state of contamination of a subject in previous time points to his/her probability of being diagnosed with PM at time $t+1$. If $c(i;t) = 0$ (and, hence, $c(i;t-k) = 0, \forall k$) subject i is not contaminated at time t , and he cannot be diagnosed at time $(t+1)$. In that case, (3) vanishes for any μ_s , $s = 1, 2, \dots, K$. If subject i is contaminated beginning in s years, $p_{mi}(t+1)$ would be equal to $\mu_1 + \mu_2 + \dots + \mu_s$, for $s \leq K$. In the case of contamination from K years (maximum latency length), $p_{mi}(t+1) = 1$. Here μ_s represents the increase in risk of being diagnosed in the transition from $s-1$ to s years from contamination. We remark that μ_s can be written as a function of π_y probabilities.

If we denote by $v(t) = \sum_i c(i;t)$, the sum w.r.t. i of Eq. (1) is equal to

$$v(t+1) = v(t) + \gamma A(t) \cdot [\text{Population at risk at time } t]. \quad (4)$$

The population at risk for contamination at a given time is formed by all residents at that time from which already contaminated subjects have to be subtracted. A sensible assumption is that

$$\text{Population at risk at time } (t) = r(t) [1 - v(t)/R(t-1)], \quad (5)$$

where $r(t)$ denotes the number of residents in the area in year t while $R(t-1)$ denotes the cumulative number of residents in the area until time $t-1$. Substitution

of (5) in (4) leads to a recurrence definition of function $v(\cdot)$:

$$v(t+1) = v(t)[1 - \gamma A(t)r(t)/R(t-1)] + \gamma A(t)r(t), \quad (6)$$

with initial value $v(1) = \gamma A(0)r(0)$.

Let us consider the aggregation of the process that describes the disease's expression. If we denote by $h(t) = \sum_i m(i;t)$, the sum w.r.t. i of Eq. (2) leads to

$$h(t+1) = h(t) + \sum_i Be(p_{mi}(t+1))I_{\{m(i;t)=0\}}.$$

If we take the expectation and substitute (3) into it, through some passages, it can be proven that $h(0) = h(1) = 0$ and, for $t \geq 2$,

$$h(t) = \sum_{s=1}^{\min(K,t-1)} \pi_s v(t-s) \quad \forall t \geq 2. \quad (7)$$

Observe that Eq. (7) could also result as a consequence of the following reasoning. The group of subjects that were contaminated at a given time $t-s$ develops the disease later (in other words, contributes to subsequent values of function $h(\cdot)$); this happens in part starting from time $t-s+1$ (with probability π_1), in part starting from time $t-s+2$ (with probability π_2), etc. up to time $t-s+K$ (with probability π_K). In a dual way, the group of diagnosed subjects at time t , $h(t)$, is composed as a mixture of subjects contaminated at different time points, and the “mixing rule” is given by the latency probabilities.

Up to this point, π_y probabilities were totally unspecific. In order to make their structure simpler and reduce the number of parameters to be estimated, a sensible assumption might be that $Y \sim 1 + Bi(K-1, p)$, where Bi denotes a Binomial distribution. This is a discrete r.v. with finite support and, as K and p vary, that distribution is flexible enough to describe different latency patterns. We remark that K represents the maximum latency length, while p does not have a specific interpretation but its value has to be read jointly with K . If we denote by $\omega(t)$ the values of the time series counting the cumulative yearly number of diagnosis, the following model results:

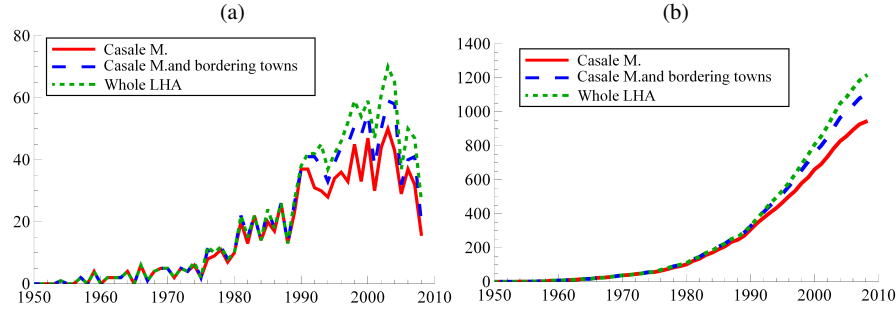
$$\omega(t) = h(t) + \varepsilon_t \quad \text{with} \quad h(t) = \sum_{s=1}^{\min(K,t-1)} \binom{K-1}{s-1} p^{s-1} (1-p)^{K-s} v(t-s), \quad (8)$$

where function $v(\cdot)$, which was defined in Eq. (6), depends upon the unknown parameter γ . Due to the complex structure of function $h(\cdot)$, in order to concretely evaluate parameters' estimates of this regression model, K was in turn kept fixed, and estimates for γ and p were evaluated conditionally based on the choice of K . Estimates were calculated with the nonlinear least squares procedure solved through the Levenberg-Marquardt algorithm. In the end, the optimal K value was chosen as the one that made the residual deviance smaller.

3 Data sources

As in other works pertaining to the same area, [4], we perform our analysis by dividing the LHA of Casale M. into three zones (the city of Casale M., the towns sharing boundaries with it, and the other towns of the LHA).

Fig. 1 Yearly time series of PM diagnosis among residents in the three zones (instantaneous (a) and cumulative (b) values).



Available data concern:

1. the number of residents from 1907 to 1986 in the three zones (source: ISTAT);
2. yearly diagnosis counts: data are obtained by integrating three sources: a) RENAM (National Mesothelioma Registry), for the period 1990–2004, b) Division of Pathological Anatomy, City Hospital, Casale M. for the period 1989–2009, and c) Public Prosecutor's office, Torino: plaintiffs' list in the proceedings to the managers of Eternit (which owned the plant).² The latter source reports diagnosis during the whole period, but the most recent years are not fully covered, since people with a recent PM diagnosis did not complete the procedure that is required for inclusion as plaintiffs yet. Because of this, the years from 2004 until 2009 are covered only by the second source. It is then plausible that our final data underestimate the true values in a non-negligible way. This entails that our predictions have to be considered only as a lower bound for future values. Finally, notice that some sources report the death year, while others report the diagnosis year. Since for this disease survival is generally shorter than one year, consistent with other studies, mortality is considered the same as the incidence;
3. yearly number of asbestos fibers processed in the plant: we used approximate values obtained through processing data about national production, the time-varying amount of that quantity produced from the plant under study, and the average asbestos quantity used for each ton of finished products.

4 Results

In Fig. 1, the yearly data of number of diagnoses among LHA residents are presented. The majority of deaths concern residents in the city of Casale M. Because of the low number of cases in the two more distant zones, we believed it was more convenient to fit a model for the first zone (Casale M.), a model for the first two zones (Casale M. and the towns sharing boundaries with it), and a third one for the whole

² The proceedings are presently ongoing, and the managers are charged with omitting prevention tools in order to reduce fiber dispersion (although the danger of cancer onset due to asbestos exposure has been known since the '60s).

LHA. The comparison among estimates of the three models allows an investigation into how distance from the plant affects the process evolution.

The parameters' estimates arising from fitting model (8) to the time series of diagnoses among residents in Casale, are shown in the second column of Table 1. The adequacy of model fitting can also be evaluated through Fig. 2. The most substantial discrepancy can be observed for $t \in [40, 65]$ (from 1947 to 1972), since the number of deaths predicted by our model is greater than those reported by our sources. This result can be explained by noticing that PM at that time was poorly known. Moreover, many workers who will have developed PM, died earlier because of asbestosis; furthermore, PM is underestimated in elderly people who rarely submit to the invasive diagnostic procedures needed to have an uncontroversial diagnosis. For this reason, in some works (see, e.g., [5]) the observed number of diagnosis is corrected through some multiplicative factors that take into account that in the past PM diagnosis were underestimated. In this work, these corrections were not applied because the possible methods do not always agree, but will be studied in the future.

According to our results, the yearly contamination probability has been for more than 30 years (from 1955 to 1985) greater than 0.0006 with a peak of 0.012 in the mid-sixties. Non-negligible probabilities are associated with latencies between 18 and 37 years, with a peak around 28/29 years ($\hat{K} = 42$ denotes the largest admissible value for Y). Among people that were residents in this city, 210 new diagnoses are expected. Starting from 2025, this phenomenon will run out.

5 Concluding remarks

Results pertaining to the three considered zones are summarized in Table 1 and suggest that, as far as residents in more distant towns from the plant are included,

- the estimated value for γ is lower, in other words, the contamination probability for an exposed subject is reduced;
- the average and the maximum value for the latency are longer;
- as a consequence of the previous statement, the year forecasted as the end of this “epidemic” is moved far away, and there is a larger number of not-yet-diagnosed cases.

Fig. 2 Observed and fitted values. City of Casale Monferrato.

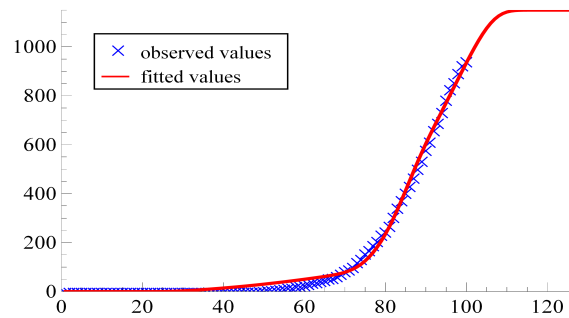


Table 1 Comparison among results pertaining to the three territories.

	Casale M.	Casale M. and bordering towns	Whole LHA
$\hat{K}$	42	49	55
$\hat{\gamma}$ (s.e.)	1.85977×10^{-8} (2.57561×10^{-10})	1.53608×10^{-8} (2.68102×10^{-10})	1.05749×10^{-8} (2.41394×10^{-10})
$\hat{\rho}$ (s.e.)	0.667041 (0.00586942)	0.623482 (0.00609547)	0.594049 (0.00684691)
R^2	0.995937	0.994283	0.991639
Peak value for the contamination probability	0.00121388	0.00100261	0.00069023
Average latency	28.35	30.93	33.08
Obs. number of diagnoses until 2009	942	1099	1211
Forecasted number of future diagnoses	208	353	479
Year when diagnoses will end	2025	2028	2033

Acknowledgements This research was supported by Fondazione CARIPARO, Project “Innovation Diffusion Processes: Differential Models, Agent-Based Frameworks and Forecasting Methods”. Progetti di Eccellenza 2007–2008.

References

- [1] Boccara, N.: Modeling Complex Systems. Springer–Verlag, New York (2004)
- [2] Guseo, R., Guidolin, M.: Modelling a dynamic market potential: A class of automata networks for diffusion of innovations. *Technol. Forecast. and Soc. Change* **76**(6), 806–820 (2009)
- [3] Hillerdal, G.: Mesothelioma: cases associated with non-occupational and low dose exposures. *Occup. Environ. Med.* **56**, 505–513 (1999)
- [4] Magnani, C., Terracini, B., Ivaldi, C., Botta, M., Mancini, A., Andron, A.: Pleural malignant mesothelioma and non-occupational exposure to asbestos in Casale Monferrato, Italy. *Occup. Environ. Med.* **52**, 362–367 (1995)
- [5] Marinaccio, A., Montanaro, F., Mastrantonio, M., Uccelli R. Altavista, P., Nesti, M., Seniori Costantini, A., Gorini, G.: Predictions of mortality from pleural mesothelioma in Italy: a model based on asbestos consumption figures supports results from age-period-cohort models. *Int. J. Cancer* **115**, 142–147 (2005)
- [6] Maule, M., Magnani, C., Dalmaso, P., Mirabelli, D., Merletti, F., Biggeri, A.: Modeling mesothelioma risk associated with environmental asbestos exposure. *Environ. Health Perspect.* **115**(7), 1066–1071 (2007)
- [7] Merler, E., Lagazio, C., A., B.: Andamento temporale della mortalità per tumore primitivo pleurico e incidenza del mesotelioma pleurico in Italia: una situazione particolarmente grave. *Epidemiol. Prev.* **23**, 316–326 (1999)
- [8] Pelucchi, C., Malvezzi, M., La Vecchia, C., Levi, F., Decarli, A., Negri, E.: The mesothelioma epidemic in western Europe: an update. *British Journal of Cancer* **90**, 1022–1024 (2004)
- [9] Peto, J., Decarli, A., La Vecchia, C., Negri, E.: The European mesothelioma epidemic. *British Journal of Cancer* **79**, 666–672 (1999)