P. Pietra

An up-wind finite element method for a filtration problem


<http://www.numdam.org/item?id=M2AN_1982__16_4_463_0>
AN UP-WIND FINITE ELEMENT METHOD FOR A FILTRATION PROBLEM (*)

by P. PIETRA (1)

Communiqué par F. BREZZI

INTRODUCTION

It is well known (see e.g. [9]) that the study of the flow of an incompressible fluid through a porous medium leads to free boundary problems for elliptic equations. These problems were initially studied with heuristic methods, applying a fixed point procedure for a sequence of problems, each of which solved on a different fixed domain (see e.g. [14], [17]). A great improvement to the theory was introduced by Baiocchi (see e.g. [4]), who formulated the problem, in the special case of a rectangular domain, on a rigorous mathematical basis, transforming it into a variational inequality of obstacle type. This idea was then generalized (see e.g. [8], [5]) for the treatment of more general domains. According to necessary, the free boundary problem was transformed into a variational inequality depending on one or more additional parameters or into a quasi variational inequality. These formulations, if applicable, are very good and give

(*) Received in December 1981

(1) Istituto di Analisi Numerica del C.N.R corso Carlo Alberto 5, 27100 Pavia, Italie.
rise to efficient numerical algorithms. However, their application requires some restriction on the geometry of the domain. Also, different geometry and boundary conditions may lead to different formulations.

Later on, new formulations were given by Brezis-Kinderleherer-Stampacchia [10] and Alt [1]. This new framework is somehow more complicated, the solution is less regular, but allows a more general treatment, in particular with respect to the geometry of the domain. The more general formulation by Alt [1] was also treated from the numerical point of view. In [2], Alt proves, under suitable assumptions on the discretization, the existence of a discrete solution and the strong convergence of a subsequence of them to a solution of the continuous problem. The abstract framework consists of conforming finite element methods and no “approximation” is done for the differential operators. This fact somehow restricts the choice of the “available” types of triangulation and all the given examples (see [2]) require decompositions of uniform type. Hence, we are back to some kind of finite differences framework and the treatment of the fixed boundaries requires some adjustment, even in the case of a rectangular domain.

In this paper, we deal with Brezis-Kinderleherer-Stampacchia's formulation, which is similar to but slightly more restrictive than Alt's one (e.g., capillarity effects are neglected). We remain, essentially, within the framework of conforming finite elements, but we introduce some up-wind techniques into the discretization. In order to justify this idea, we note that the pressure $u$ verifies an equation of type

$$ - \Delta u - D_y H(u) = 0 \quad \text{in } \Omega, $$

where $H(x)$ is the Heaviside function, so that $H(u)$ is the characteristic function of the set $\{ u > 0 \}$.

Equation (0) can be interpreted as a diffusion-convection problem, where the coefficient of the convection term may be infinite. It is well known that usual finite element methods are not suitable for this kind of problems. In fact, these methods are unstable, when the ratio between the diffusion coefficient and the convection coefficient is « too small ». Therefore, some special « up-wind treatment of the convection term (essentially, of the $D_y$ operator) has to be used in order to recover stability. Many of these up-wind techniques are known in the literature for finite element methods (see e.g. [18] and the references therein contained). Here we choose a scheme introduced by Tabata [19].

With respect to Alt's scheme, our approximation has the disadvantages that the formulation (as previously stated) is more restrictive. Moreover, only a weak convergence can be proved. On the other hand, many arguments are much simpler (also because the problem is not considered in an abstract setting).
Nevertheless the treatments are similar enough, so that many crucial Alt's arguments may be used. The present formulation allows much more general decompositions, so that we are truly in the framework of finite element methods, and therefore the fixed boundary is "followed" much more neatly. For instance each polygon can be treated without approximation of the domain and automatic decomposition routines can be used. Hence the formulation is more suitable for entering a finite element code.

What follows is an outline of the paper: in paragraphs 1 and 2 we state respectively the continuous and the approximate problem; in paragraph 3 we prove the existence of a discrete solution and in paragraph 4 we prove that a subsequence of discrete solutions converges to a solution of the continuous problem; paragraph 5 reports some numerical results and finally paragraph 6 contains some concluding remarks.

In this paper we will use the classical Sobolev spaces with the following notations:

\[ \| v \|_{k, \Omega} = \left( \int_{\Omega} \sum_{|\alpha| \leq k} | \partial^\alpha v |^2 \, dx \right)^{1/2} \]

\[ | v |_{k, \Omega} = \left( \int_{\Omega} \sum_{|\alpha| = k} | \partial^\alpha v |^2 \, dx \right)^{1/2}. \]

1. THE CONTINUOUS PROBLEM

Let \( \Omega \) be the section of a porous medium. For simplicity's sake, it is assumed to be a polygon. We remark that with similar arguments it is possible to consider problems where \( \Omega \) is a bounded, connected open set of \( \mathbb{R}^2 \), with a Lipschitz boundary \( \partial \Omega \).

We denote by \( S^+ \) the part of the boundary in contact with the reservoirs, and by \( S^0 \) the part in contact with the air. The third part, \( \partial \Omega \setminus (S^+ \cup S^0) \), is the impervious part of the dam. Moreover \( S^0 \) and \( S^+ \) are measurable and disjoint sets, and the measure of \( S^+ \) is positive. The medium is assumed to be inhomogeneous and anisotropic. The permeability is given by a symmetric tensor \( K \), such that

\[ K \in (C^{0,1}(\Omega))^4 \quad \xi_i K_{ij} \xi_j \geq \alpha \, | \xi |^2 \quad \forall \xi \in \mathbb{R}^2 \quad (1.1) \]

where \( C^{0,1}(\Omega) \) is the space of the Lipschitz continuous functions.

Let \( e = (0, 1) \) be the vertical unit vector.

\(^{1}\) The convention of summation of repeated indices is assumed, and \( | \, | \) denotes here the euclidean norm.

vol. 16, n° 4, 1982
We suppose that the atmospheric pressure is zero, and we neglect the capillarity and evaporation effects.

The function \( u_0 \in C^0(S^0 \cup S^+) \) denotes the boundary value of the pressure, i.e. \( u_0 \) is the hydrostatic pressure on \( S^+(u_0 > 0 \text{ on } S^+) \), and it is zero on \( S^0 \).

We consider the following continuous problem:

\[
\begin{align*}
\text{Problem 1:} \quad & \text{Find a pair } (u, \gamma) \in H^1(\Omega) \times L^\infty(\Omega) \text{ such that } \\
& u \geq 0 \text{ a.e. in } \Omega, \ u = u_0 \text{ on } S^+ \cup S^0 \\
& 0 \leq \gamma \leq 1 \text{ a.e. in } \Omega, \ \gamma = 1 \text{ a.e. on } \{u > 0\} \\
& \int_{\Omega} \nabla \varphi \mathbf{K} (\nabla u + \gamma \varepsilon) \, dx \geq 0 \quad \forall \varphi \in W = \{w \in H^1 \mid w = 0 \text{ on } S^+, \ w \leq 0 \text{ on } S^0\}.
\end{align*}
\]

For a theoretical study of problem (1.2) we refer to the works by Brézis-Kinderlehrer-Stampacchia [10] (existence results); Alt [1] (existence and regularity results for a more general problem that, for a suitable choice of test functions, reduces to problem 1), Alt-Gilardi [3] and Chipot [12] (uniqueness results, and characterizations of non-uniqueness situations).

In case of more restrictive assumptions on the geometry of the domain \( \Omega \), other formulations of the problem are known: see, for instance, Baiocchi [4] (transformation of the problem in a variational inequality), Baiocchi [5] (transformation in a quasi-variational inequality) and Baiocchi-Capelo [6] (for complete references about these problems). These formulations are the starting point for a numerical study of the problem, see for instance [7] and [6] (for the further references).

2. THE DISCRETE PROBLEM

Let \( \{\mathcal{G}_h\}_h \) be a family of triangulations of \( \Omega \), depending on a parameter \( h > 0 \). For each triangulation \( \mathcal{G}_h = \{T_i\}_{i=1}^N \), and for each \( T_i \in \mathcal{G}_h \), we set the following notations:

\[
\begin{align*}
h(T_i) &= \text{the diameter of } T_i, \\
\rho(T_i) &= \text{the supremum of the diameters of the balls contained in } T_i, \\
h &= \max \{ h(T_i) \mid T_i \in \mathcal{G}_h \}
\end{align*}
\]

We suppose that the triangulation \( \mathcal{G}_h \) is regular and of (weakly) acute type, i.e. there exists a constant \( \sigma < 1 \), independent of the triangulation, such that

\[
h(T_i) \leq \sigma \rho(T_i), \quad \forall T_i \in \mathcal{G}_h,
\]

\( \text{R A I R O Analyse numerique/Numerical Analysis} \)
and every angle $\theta$ of the triangles of $\mathcal{C}_h$ verifies

$$\theta \leq \pi/2.$$  

(2.2)

Let $\{ P_t \}_{t \in \mathcal{N}_h}$ be the set of the nodal points of the triangulation and let us consider some subsets of the index set $\mathcal{N}_h$:

$$N_{h}^+ = \{ i \in \mathcal{N}_h \mid P_t \in S^+ \}; \quad N_{h}^0 = \{ i \in \mathcal{N}_h \mid P_t \in S^0 \};$$

$$N_h = \mathcal{N}_h \setminus \mathcal{N}_h; \quad \mathcal{N}_h = \{ i \in \mathcal{N}_h \mid P_t \notin \partial \Omega \text{ or } K(P_t) \subseteq \text{applied at } P, \text{ intersects } \Omega \} .$$

We also introduce a dual decomposition of the domain $\Omega$:

$$\overline{\Omega} = \bigcup_{i \in \mathcal{N}_h} D_i,$$

(2.3)

where $D_i$ is the barycentric domain associated with $P_t$, i.e.

$$D_i = \bigcup_k \{ D_i^k \mid T_k \in \mathcal{C}_h \text{ s.t. } P_t \text{ is a vertex of } T_k \} ,$$

(2.4)

where

$$D_i^k = \bigcap_{j=1,2} \{ x \mid x \in T_k, \lambda_j^i(x) \leq \lambda_i(x) \} ,$$

and $\lambda_0, \lambda_1, \lambda_2$ are the barycentric coordinates with respect to the vertices of $T_k, P_t, P_t^1, P_t^2$ (see fig. 1).

Moreover let $\Omega_h$ be defined as:

$$\Omega_h = \bigcup_{i \in \mathcal{N}_h} D_i^0 .$$

We choose

$$V_h = \{ v_h \in C^0(\overline{\Omega}), v_h|_{T_t} \in P_1, \forall T_t \in \mathcal{C}_h \} ,$$

(2.5)

where $P_1$ denotes the space of polynomials of degree $\leq 1$; we call $\phi_i^h$ the basis function of $V_h$ associated with the nodal point $P_t$, i.e. such that $\phi_i^h(P_j) = \delta_{ij}$.
Let \( \chi_h^i \) be the characteristic function of the domain \( D_i \), and let \( \Psi_h \) be the linear space spanned by \( \chi_h^i, i \in \mathcal{N}_h \); i.e.
\[
\Psi_h = \{ \psi_h \in L^2(\Omega) \mid \psi_h(x) = \sum_{i \in \mathcal{N}_h} \psi_h^i \chi_h^i(x) \}.
\] (2.6)

Finally, we introduce the up-wind triangle associated with the nodal point \( P_i \) (see [19]).
A triangle \( \mathcal{U}_i \subseteq \mathcal{C}_h \) is called the up-wind triangle of the nodal point \( P_i \) if :

(i) \( P_i \) is a vertex of \( \mathcal{U}_i \),

(ii) \( \mathcal{U}_i \setminus P_i \) intersects the oriented half-line with end point \( P_i \) and direction \( K(P_i) \in (\text{see fig. 2 for } K = kI). \)

![Figure 2.](image_url)

We remark that all the nodal points \( P_i \) with \( i \in \mathcal{N}_h \), have an up-wind element. If there exist two up-wind elements at the node \( P_i \), we call \( \mathcal{U}_i \) one of them at arbitrarily chosen.

We now define a linear operator \( E_h \) from \( V_h \) into \( \Psi_h \) in the following way :
\[
E_h w_h = \sum_{i \in \mathcal{N}_h} (E_h, w_h) \chi_h^i, \text{ for each } w_h \in V_h,
\] (2.7)

where
\[
E_h, w_h = K(P_i) \psi(\nabla w_h|\mathcal{U}_i),
\] (2.8)
i.e. \( E_h w_h \) is a function of \( \Psi_h \) such that \( E_h w_h(P_i) \) is the value of the derivative in the \( K(P_i) \) direction of the function \( w_h \) on the triangle \( \mathcal{U}_i \). We remark that this derivative is a constant in \( \mathcal{U}_i \) and, if there are two up-wind elements, \( E_h w_h \) is independent of the choice of \( \mathcal{U}_i \), since \( w_h \in V_h \).

We denote by \( \tilde{u}_0 \) the value of \( u_0 \) at \( P_i \).
The following discretized problem can now be introduced:

**Problem 2**: Find a pair \((u_h, \gamma_h) \in V_h \times \Psi_h\) such that

\[
\begin{align*}
(i) & \quad u_h \geq 0 \text{ in } \Omega \text{ and } u'_h = u'_0 \text{ if } i \in N_h^+ \setminus N_h^0 \\
(ii) & \quad 0 \leq \gamma_h \leq 1 \text{ in } \Omega \text{ and } \gamma'_h = 1 \text{ if } u'_h > 0 \\
(iii) & \quad a_h(u_h, v_h) + (E_h v_h, \gamma_h) \geq 0
\end{align*}
\]

\(\forall v_h \in W_h = \{ w_h \in V_h \mid w'_h = 0 \text{ if } i \in N_h^+, w'_h \leq 0 \text{ if } i \in N_h^0 \} \),

where \((\cdot, \cdot)_h\) denotes the scalar product in \(L^2(\Omega_h)\) and

\[a_h(u_h, v_h) = \int_{\Omega} \nabla u_h \cdot K_h \nabla v_h \, dx\]

with \(K_h\) constant on each triangle, defined by \(K_h(x) = K(b_T) \forall x \in T\), where \(b_T\) is the barycentre.

3. **Existence Results**

In order to prove the existence of a solution of the discrete problem (2.9), a different form is adapted.

Defining \(a_{ij}\) and \(e_{ij}\) as follows

\[a_{ij} = a_h(\phi^i_h, \phi^j_h) \quad i, j \in N_h \]

\[e_{ij} = \begin{cases} (E_h \phi^i_h, \gamma^j_h)_h & \text{if } j \in \overline{N_h} \\ 0 & \text{if } j \in N_h \setminus \overline{N_h} \end{cases} \]

we can verify that the inequality (2.9) is equivalent to

\[\sum_{j \in \overline{N_h}} (a_{ij} u'_h + e_{ij} \gamma'_h) = 0 \quad \text{if } i \in N_h \setminus (N_h^+ \cup N_h^0) \]

\[\leq 0 \quad \text{if } i \in N_h^0 \]

\[u'_h = u'_0 \quad \text{and} \quad \gamma'_h = 1 \quad \text{if } i \in N_h^+ .\]

A theorem, the proof of which can be found in Alt [2] (theorem 2.4), is stated hereafter.

**Theorem 3.1**: Let \(a_{ij}\) and \(e_{ij}\) be defined as in (3.1) and (3.2). If

\[a_{ii} > 0; \quad a_{ij} \leq 0 \quad \text{if } j \neq i \]

\[e_{ii} \geq 0; \quad e_{ij} \leq 0 \quad \text{if } j \neq i ,\]

then there exists a pair \((u_h, \gamma_h) \in V_h \times \Psi_h\) solution of problem (2.9).
Hence it is sufficient to show that the hypotheses (3.4) and (3.5) are verified in our case.

**LEMMA 3.1:** Let $a_{ij}$ and $e_{ij}$ be defined by (3.1) and (3.2).

Let $\phi_T(K)$ be defined by

$$\cos \phi_T = \frac{2}{\sqrt{2 + \lambda + 1/\lambda}}, \quad 0 \leq \phi \leq \pi/2,$$

where $\lambda = \| K(b_T)^{-1} \| \| K(b_T) \|$.

If $\mathcal{C}_h$ is such that for each triangle $T \in \mathcal{C}_h$ and for each angle $\theta$ in $T$ we have

$$\theta \leq \pi/2 - \phi_T,$$

(3.6)

then the following inequalities hold:

(i) $a_{ii} > 0; a_{ij} \leq 0$ if $j \neq i$

(ii) $e_{ii} \geq 0; e_{ij} \leq 0$ if $j \neq i$.

**Proof:** In the isotropic case ($\lambda = 1$), it is well known (see e.g. [13]) that the assumptions on the decomposition (2.1), (2.2), and the property (1.1) on $K$ imply (i). By analogous arguments we can show that (3.6) implies (i). We remark that it is possible to consider problems with weak anisotropy. In particular $\theta$ is equal to $\pi/3$ if $\lambda = 3$.

The proof of (ii) is contained in Tabata [19] (lemma 3), but we recall it for reader’s convenience.

Let $P_i, P_{j_1}, P_{j_2}$ be the vertices of $\mathcal{C}_h$ and $\lambda_p, \lambda_{j_1}, \lambda_{j_2}$ its barycentric coordinates.

We remark that

$$e_{ij} = \begin{cases} (\chi_{h}, \chi_{k}) K(P_i) e \nabla \lambda_i & \text{if } i \in \{ j, j_1, j_2 \} \\ 0 & \text{otherwise} \end{cases}$$

Let $\bar{p}_k$ be the vector $\overrightarrow{P_i P_{j_k}}, k = 1, 2$.

A short calculation shows that

$$\nabla \lambda_{j_k} \cdot \bar{p}_k = \delta_{lk} \quad \text{for } l, k = 1, 2 \quad (3.7)$$

$$\nabla \lambda_{j_k} \cdot \bar{p}_k = -1 \quad \text{for } k = 1, 2. \quad (3.8)$$

By the definition of up-wind triangle, it follows that there exist non-negative numbers $c'_{k}, k = 1, 2$, such that

$$K(P_i) e = - c'_{1} \bar{p}_1 - c'_{2} \bar{p}_2. \quad (3.9)$$

From (3.7), (3.8) and (3.9), (ii) can be obtained.
Remark 3.1 We note explicitly that in the isotropic case \( K = kI \) one has \( \phi_T = 0 \) and condition (3.6) reduces to (2.2).

Remark 3.2 For the actual computation of a solution \((u_h, \gamma_h)\) of a problem of this type, we refer to [2].

4 CONVERGENCE RESULTS

In this section it will be proved that it is possible to extract a subsequence that converges to a solution of the continuous problem, from each family of discrete solutions. More precisely we will prove the following result.

**Theorem 4.1** For each family \( \{ (u_h, \gamma_h) \} \) of solutions of problem (2.9), there exists a subsequence \( \{ (u_{h_k}, \gamma_{h_k}) \}_{k=0}^{\infty} \), and there exists a pair

\[
(u, \gamma) \in H^1(\Omega) \times L^\infty(\Omega)
\]

such that

\[
u_{h_k} \rightharpoonup u \text{ weakly in } H^1(\Omega)
\]

\[
\gamma_{h_k} \rightharpoonup^* \gamma \text{ weakly star in } L^\infty(\Omega)
\]

The pair \((u, \gamma)\) is a solution of problem (1.2).

If \((u, \gamma)\) is the unique solution of problem (1.2), the whole sequence \( \{ (u_h, \gamma_h) \} \) converges to \((u, \gamma)\).

In order to prove theorem 4.1, some lemmas are needed. We define for each \( v \in H^1(\Omega) \)

\[
E_h v = \sum_{i \in \mathcal{N}_h} \chi_h \left( \int_{v_{P_i}} (K(P_i) \in \nabla v) \, dx \right), \tag{4.1}
\]

where

\[
\int_{v} f \, dx = \frac{1}{m(T)} \int_T f \, dx
\]

We remark that if \( v \in V_h \), the definition (4.1) coincides with the previous definition (2.7).

**Lemma 4.1** There exists a constant \( c \) independent of \( h \), such that

\[
| (E_h v, \eta)_h | \leq c \| v \|_{1, \Omega} \| \eta \|_{0, \Omega}, \tag{4.2}
\]

for each \( v \in H^1(\Omega) \) and for each \( \eta \in L^2(\Omega) \).

**Proof** From Holder's inequality it follows that

\[
| (E_h v, \eta)_h | \leq c \| E_h v \|_{0, \Omega} \| \eta \|_{0, \Omega}
\]
Moreover
\[ \| E_h v \|_{0, \Omega_h}^2 = \int_{\Omega_h} \left( \sum_{i \in \mathcal{N}_h} \left( \int_{\mathcal{T}_i} K(P_i) e \nabla v \, dx \right) \chi_h^i \right)^2 = \sum_{i \in \mathcal{N}_h} \left\| \int_{\mathcal{T}_i} K(P_i) e \nabla v \, dx \right\|_{0, \Omega_i}^2 \]

On the other hand
\[ \left\| \int_{\mathcal{T}_i} K(P_i) e \nabla v \, dx \right\|_{0, \Omega_i}^2 \leq \frac{m(D_i)}{m(\mathcal{N}_i)} \left\| K(P_i) e \nabla v \right\|_{0, \mathcal{T}_i}^2 \leq c \| \nabla v \|_{0, \mathcal{T}_i}^2, \]

because of the previous assumption of regular triangulation.
Therefore we have
\[ \| E_h v \|_{0, \Omega_h}^2 \leq c \| v \|_{1, \Omega}^2 , \]
and (4.2) is proved.

**Lemma 4.2:** For each \( v \in H^1(\Omega) \), for each \( \eta \in L^2(\Omega) \)
\[ | (E_h v, \eta)_h - (Ev, \eta) | \to 0 \quad \text{for } h \text{ vanishing}, \quad (4.3) \]
where \( (Ev, \eta) = \int_{\Omega} (Ke \nabla v) \eta \, dx \).

**Proof:**
\[ | (E_h v, \eta)_h - (Ev, \eta) | \leq | (E_h v, \eta)_h - (Ev, \eta)_h | + \left| \int_{\Omega \setminus \Omega_h} Ke \nabla v \eta \, dx \right| . \]

The second term goes to zero since
\[ m(\Omega \setminus \Omega_h) \to 0 \quad \text{when } h \to 0 . \]

In order to prove that also the first term vanishes, a first step is to show that for each \( w \in C^\infty(\Omega) \) and each \( \eta \in L^2(\Omega) \)
\[ | (E_h w, \eta)_h - (Ew, \eta)_h | \leq c h \quad (4.4) \]
holds.

We have
\[ | (E_h w, \eta)_h - (Ew, \eta)_h | \leq | (E_h w, \eta)_h - (I_h w, \eta)_h | + | (I_h w, \eta)_h - (Ew, \eta)_h | , \]
where
\[ I_h w = \sum_{i \in \mathcal{N}_h} (I_{h,i} w) \chi_i, \quad \text{with} \quad I_{h,i} w = K(P_i) \nabla w. \]

By Hölder's inequality and by assumption (1.1) on \( K \), we obtain
\[ |(E_h w, \eta)_h - (E w, \eta)_h| \leq \|E_h w - I_h w\|_{0, \Omega_h} \|\eta\|_{0, \Omega_h} + ch |w|_{1, \Omega_h} \|\eta\|_{0, \Omega_h}. \]

Moreover
\[ \|E_h w - I_h w\|_{0, \Omega_h}^2 = \sum_{i \in \mathcal{N}_h} \|E_{h,i} w - I_{h,i} w\|_{0, \Omega_i}^2. \]

Now the problem is to obtain an uniform estimate of \( \|E_{h,i} w - I_{h,i} w\|_{0, \Omega_i} \)
\( w \in C^\infty(\Omega) \), then there exists a point \( \xi_i \in \mathcal{U}_i \) such that
\[ K(P_i) \nabla w(\xi_i) = \int_{\mathcal{U}_i} K(P_i) \nabla w \, dx. \]

From this remark and recalling in particular that \( \nabla w \) is a Lipschitz continuous function, we obtain
\[ \|E_{h,i} w - I_{h,i} w\|_{0, \Omega_i}^2 = \int_{\mathcal{U}_i} \frac{|K(P_i) \nabla w(\xi_i) - K(P_i) \nabla w(x)|^2}{|\xi_i - x|^2} |\xi_i - x|^2 \, dx \leq ch^2 m(D_i). \]

Therefore
\[ |(E_h w, \eta)_h - (I_h w, \eta)_h| \leq ch \]
holds, and (4.4) follows.

In order to complete the proof, we recall that \( C^\infty(\Omega) \subset H^1(\Omega) \) with density; so for each \( v \in H^1(\Omega) \), there exists a sequence \( \{w_n\} \) of functions \( w_n \in C^\infty(\Omega) \)
\( w_n \to w \) strongly in \( H^1(\Omega) \). Then by (4.2) and (4.4), applied to \( w_n \in C^\infty(\Omega) \), (4.3) easily follows.

**Lemma 4.3** : If the triangulation \( \mathcal{T}_h \) verifies (2.1), (2.2), and \( a_{ij}, e_{ij} \) are defined as in (3.1) and (3.2), then the following properties hold:

(i) there exists a constant \( \beta_1 > 0 \), such that
\[ e_{ii} \leq \beta_1 \eta a_{ii} \text{ for each } i \in \mathcal{N}_h, \beta_1 \text{ independent of } \mathcal{T}_h \]

(ii) if \( P_k \) and \( P_j \) are nodal points such that there exists a constant \( \beta_2 \) with
\[ a_{kk} \leq -\beta_2 a_{kj}, \]

vol 16, no 4, 1982
then
\[ u'_h \leq \beta (u^k_h + h), \]
where \( \beta = \max (\beta_2, \beta_1 \beta_2). \)

**Proof**: Property (i) easily follows by assumptions (3.1) and (3.2).
Since \( \gamma^k_h \leq 1 \), from the inequality (3.3)
\[ a_{kk} u^k_h + e_{kk} \geq - a_{kj} u'_h, \]
follows, hence
\[ u'_h \leq \frac{a_{kk}}{a_{kj}} u^k_h + \frac{e_{kk}}{-a_{kj}}. \]

By assumptions
\[ a_{kk} \leq - \beta_2 a_{kj}, \]
moreover
\[ e_{kk} \leq \beta_1 h a_{kk}, \]
then
\[ u'_h \leq \beta (u^k_h + h), \]
where \( \beta = \max (\beta_2, \beta_1 \beta_2). \)

**Lemma 4.4**: Let \( i \in N_h \) be an index such that \( \gamma^i_h < 1 \).
Then there exists a constant \( C \), independent of the triangulation \( T_h \), such that
\[ u_h \leq C h \text{ on } D_i \]
(4.6)
holds.

**Proof**: Let \( J^i_h \) be the set of indices \( j \) such that the node \( P_j \) is adjacent to \( P_i \).
We have to show that
\[ u^i_h \leq C h \]
(4.7)
for each \( j \in J^i_h \).
The condition \( \gamma^i_h < 1 \) together with (2.9) (ii) implies
\[ u^i_h = 0; \]
(4.8)
hence (4.7) holds trivially for \( j = i \).

The proof of (4.7) will be carried out using (4.8) and lemma 4.3(ii). The difficulty is that (4.5) is not verified for each \( j \in J^i_h \) with \( \beta_2 \) independent of \( T_h \).
We now consider a triangle $T$ that contains $P_v$. We denote by $P_{j_1}$, $P_{j_2}$, $P_{j_3}$ the vertices of $T$ and introduce the $3 \times 3$ matrix

$$a^T_{rs} = \int_T \nabla \lambda_r K_h \nabla \lambda_s \, dx \quad r, s = 1, 3$$

where $\lambda_k$ is the barycentric coordinate with $\lambda_k(P_{j_1}) = \delta_{kl}$.

It is necessary to show that there exists a constant $\beta_2 > 0$, independent of $T$ and $\mathcal{C}_h$, such that

$$\sup_r a^T_{rr} \leq - \beta_2 a^T_{rs}$$

for at least two off diagonal elements $a^T_{rs}$.

Since $\sum_{r=1}^{3} \lambda_r = 1$, we get

$$\sum_{r=1}^{3} a^T_{rs} = 0 \quad s = 1, 3.$$ (4.10)

On the other hand, with our assumptions we have

$$a^T_{rr} > 0 \quad \text{and} \quad a^T_{rs} \leq 0 \quad r \neq s \quad (\text{see lemma 3.1(i))}.$$ (4.11)

Moreover the triangulation is regular, hence

$$c_1 \leq a^T_{rr}/a^T_{ss} \leq c_2 \quad r, s = 1, 3.$$ (4.12)

Then (4.10) together with (4.11) and (4.12) implies (4.9).

Now it is easy to see, using (4.9) and the fact that the decomposition is regular, that there exists a constant $\bar{\beta}_2$, independent of $h$, such that for any node $P_v$ and any $j \in J_h^i$, at least one of the following two properties holds:

(a) $a_{ii} \leq - \bar{\beta}_2 a_{ij}$
(b) $\exists k \in J_h^i$ s.t. $a_{ii} \leq - \bar{\beta}_2 a_{ik}$ and $a_{kk} \leq - \bar{\beta}_2 a_{kk}$.

We can now conclude the proof. Let $P_v$ again be such that $\gamma_h^i < 1$ (and hence $u_h^i = 0$) and let $j \in J_h^i$. If (4.13) (a) holds, then $u_h^i \leq C(u_h^i + h)$ thanks to lemma 4.3(ii). If (4.13) (b) holds, then $u_h^i \leq C(u_h^i + h)$ and $u_h^i \leq C(u_h^i + h)$ using twice 4.3(ii). Hence (4.7) is proved.
Proof of theorem 4.1: By lemma 4.1 and the fact that $0 \leq \gamma_h \leq 1$ for each $i$, we obtain
\[
\| \gamma_h \|_{L^\infty(\Omega)} \leq 1, \\
\| u_h \|_{1,\Omega} \leq C,
\]
where $C$ is independent of $h$.

Then there exist $\gamma \in L^\infty(\Omega)$, $u \in H^1(\Omega)$ and there exist a subsequence of \{ $\gamma_h$ \} and a subsequence of \{ $u_h$ \}, that is still denoted by \{ $(u_h, \gamma_h)$ \}, such that
\[
\gamma_h \overset{*}{\rightharpoonup} \gamma \quad \text{weakly star in } L^\infty(\Omega) \\
u_h \rightharpoonup u \quad \text{weakly in } H^1(\Omega).
\]
(4.14) (4.15)
Moreover $0 < \gamma < 1$ and $u \geq 0$ almost everywhere, and $u = u_0$ on $S^0 \cup S^+$. 

Now it is necessary to prove that this pair $(u, \gamma)$ is a solution of problem (1.2). It is well known (see [16]) that for each $v \in W$ it is possible to choose a sequence of $v_h \in W_h$ such that
\[
v_h \rightarrow v \quad \text{strongly in } H^1(\Omega).
\]
(4.16)

Letting $h \rightarrow 0$ in the inequality (2.9) (iii), by (4.14), (4.15), (4.16) and lemma 4.2, we obtain
\[
\int_{\Omega} \nabla v K(\nabla u + \gamma g) \, dx \leq 0 \quad \forall v \in W. 
\]
(4.17)

In order to conclude that the pair $(u, \gamma)$ is a solution of problem (1.2), we have to show that there exists a set $N \subseteq \Omega$ of measure zero, such that
\[
\{ \gamma < 1 \} \setminus N \subseteq \{ u = 0 \} \setminus N.
\]
(4.18)

The property (4.18) is proved in [2] (theorem 3.4), but for completeness we shall give the proof below.

Let $\varepsilon > 0$ and $x$ be a point such that the set $\{ \gamma \leq 1 - \varepsilon \}$ has density 1 at $x$, i.e.
\[
\int_{B_r(x)} \chi(\{ \gamma \leq 1 - \varepsilon \}) \geq 1 - \lambda(r), \quad \text{with } \lambda(r) \rightarrow 0 \text{ when } r \rightarrow 0,
\]
where $B_r(x)$ denotes the ball of radius $r$ and centre $x$, and $\chi$ the characteristic function.
The property (4.18) is proved when we show that the set \( \{ u = 0 \} \) has lower density positive at \( x \).

Since \( \gamma \leq 1 \),
\[
\int_{B_r(x)} \gamma \leq 1 - \epsilon + \int_{B_r(x)} \chi(\{ \gamma > 1 - \epsilon \}) \leq 1 - \epsilon + \lambda(r) \leq 1 - 3/4 \epsilon
\]
holds, for \( r \) small enough.

As \( \gamma_h \rightharpoonup \gamma \) weakly in \( L^1(\Omega) \), for fixed \( r \) and \( h \) small enough, we have
\[
\int_{B_r(x)} \gamma_h \leq 1 - \epsilon/2 ,
\]
and
\[
\int_{B_r(x)} \chi(\{ \gamma_h < 1 \}) = 1 - \int_{B_r(x)} \chi(\{ \gamma_h = 1 \}) \geq 1 - \int_{B_r(x)} \gamma_h \geq \epsilon/2 . \tag{4.19}
\]

Therefore, there exists \( i \in N_h \) such that
\[
\gamma'_h < 1 \quad \text{and} \quad D_i \cap B_r(x) \neq \emptyset . \tag{4.20}
\]
By lemma 4.4
\[
u_h \leq C h \ \text{in} \ D_i \tag{4.21}
\]
holds.

Since (4.21) is true for each \( i \in N_h \) with property (4.20), we conclude that
\[
\{ \gamma_h < 1 \} \cap B_r(x) \subset \{ u_h \leq Ch \} ;
\]
and, by (4.19), that
\[
\epsilon/2 \leq \int_{B_r(x)} \chi(\{ \gamma_h < 1 \}) \leq \int_{B_r(x)} \chi(\{ u_h \leq Ch \}) .
\]

For \( \delta > 0 \) with \( h \leq \delta/C \), we introduce the function
\[
\xi_h = \max(\min(2 - u_h/\delta, 1), 0) .
\]
Hence
\[
\xi_h \geq \chi(\{ u_h \leq Ch \}) ,
\]
and
\[
\epsilon/2 \leq \int_{B_r(x)} \xi_h .
\]
Since \( u_h \to u \) strongly in \( L^1(\Omega) \)

\[
\int_{B_r(x)} \xi_h \to \int_{B_r(x)} \max \left( \min \left( 2 - u/\delta, 1 \right), 0 \right) \leq \int_{B_r(x)} \chi\left( \{ u < 2\delta \} \right).
\]

For Beppo Levi's theorem, letting \( h \to 0 \) we obtain

\[
\int_{B_r(x)} \chi\left( \{ u = 0 \} \right) \geq \varepsilon/2
\]

for \( r \) small enough, i.e. the lower density of the set \( \{ u = 0 \} \) at point \( x \) is positive. In this way, property (4.18) is proved.

5. NUMERICAL RESULTS

In order to obtain information on the accuracy of the proposed method, we tested the discrete scheme in a simple case and we compared the obtained results with the "exact solution".

The dam was supposed to be rectangular, the medium homogeneous and isotropic. We choose as "exact solution" the solution of the same problem computed via Baiocchi's transform (see [4]) with a mesh size \( h = 1/60 \). The transformation leads to the resolution of a variational inequality in the new unknown \( w \), with \( -w_y = u \). If the space \( H^1(\Omega) \) is approximated by piecewise linear finite elements, it can be proved (see [11], [15]) that the following error estimate

\[
\| w - w_h \|_{1,\Omega} \leq ch
\]

holds. Hence the choice of \( -w_{h,y} \) as "exact solution" is reasonable. We computed the relative error in \( L^2 \) and \( H^1 \) norm:

\[
ERR_i(h) = \frac{\| u_h - (-w_{h,y}) \|_{i,\Omega}}{\| w_{h,y} \|_{i,\Omega}} \quad i = 0, 1.
\]

The obtained results are as follows:

\[
\begin{align*}
h = 1/10 & \quad ERR_0(h) = 0.0019 \quad ERR_1(h) = 0.071 \\
h = 1/15 & \quad ERR_0(h) = 0.0012 \quad ERR_1(h) = 0.053 \\
h = 1/20 & \quad ERR_0(h) = 0.0009 \quad ERR_1(h) = 0.047.
\end{align*}
\]

Via least squares, we computed the convergence rate in both cases, i.e. the numbers \( a_i \) such that

\[
ERR_i(h) \leq ch^{a_i} \quad i = 0, 1
\]
and we obtained

\[ \alpha_0 = 1.08 \quad \alpha_1 = 0.59. \]

In figure 3 we reported the free boundary of the "exact solution" and the characteristic functions of the set \( \{ u_h > 0 \} \) \(^2\) for \( h = 1/10, h = 1/15, \) \( h = 1/20. \)

The numerical computations were carried out on the Honeywell 6040 system of the Centro di Calcoli Numerici of the University of Pavia.

6. CONCLUDING REMARKS

We summarize here, for simplicity's sake, the results obtained in the case of isotropic homogeneous materials (i.e. \( K = I \)) for the problem (1.2). If: \( a) \) the triangulation is of weakly acute type (see condition (2.2)); \( b) \) \( u_h \) is assumed to be piecewise linear; \( c) \) \( y_h \) piecewise constant on the dual decomposition (2.3), and \( d) \) the up-wind scheme (2.7) is chosen for the discretization of the \( D_s \) operator, then for each \( h > 0 \) the discrete problem (2.9) has at least one solution; moreover from each family \( \{ (u_h, y_h) \}_{h>0} \) of solutions we can extract a subsequence which converges weakly to a solution \( (u, y) \) of (1.2). Obviously if problem (1.2) is known to have a unique solution, then the whole sequence \( \{ (u_h, y_h) \} \) converges weakly to it. The method adapts immediately to any polygonal domain \( \Omega \) with no changes in the geometry. The implementation is reasonably simple and proved to give satisfactory numerical results.

\(^2\) Even if in the continuous problem \( \gamma = \chi(\{ u > 0 \}), \gamma_h \) is not a characteristic function, but there exists a strip of \( h \) width, with \( 0 < \gamma_h < 1. \)
AN UP-WIND METHOD FOR A FILTRATION PROBLEM

REFERENCES

[18] P A Raviart, Approximation numerique des phenomenes de diffusion-convection (1979), Ecole d'ete d'analyse numerique, CEA, EDF, IRA

vol 16, n° 4, 1982